PP 价格月度预测

2025-2026 学年《数据挖掘与机器学习》课程个人 Big Project

姓名　(学号)

2026 年 1 月 11 日

（请将\StudentName 与\StudentID 替换为真实姓名与学号；最终提交 PDF 文件名按作业要求重命名）

目录

1 问题背景与任务定义

聚丙烯（PP）是典型的大宗化工品，价格既受上游原料与能源成本（丙烯、乙烯、煤/甲醇等路线）驱动，也受供给端（产量、进口、检修）与需求端（下游开工率）影响，并通过库存与期货预期反映行业周期。企业在日常经营中需要对未来 1 个月的 PP 价格水平、涨跌方向与波动强弱做出判断，以辅助采购、排产、库存与风险对冲决策。

本项目使用企业提供的 `PP 数据/`（表 0–16）作为原始数据源：目标为华东市场 PP 粒现货价格（表 1），特征覆盖成本、供需、库存、期货与宏观等多维指标（表 2–16）。由于各表频率（日/周/月/年）与起止时间差异显著，我们需要先完成统一月度化与时序对齐，再开展建模与评估。

根据题目要求，本报告依次解决：

- **问题 1：**月度预测 PP 价格，并输出未来涨跌方向（涨/跌）。
- **问题 2：**预测涨跌强度（五分类示例：大涨/小涨/平/小跌/大跌），并输出涨跌幅度的数值预测与概率（`predict_proba`）。
- **问题 3：**考虑行业周期分阶段，设计自动筛选关键因子组合的方法，并给出各阶段因子影响权重。

表 1: 三问任务定义、输出形式与评价指标概览

任务	输出/目标	主要指标
问题 1	回归：预测下一月价格 y_t；分类：输出方向 $\mathbb{I}(y_t - y_{t-1} > 0)$	MAE/RMSE/MAPE；Acc/-Precision/Recall
问题 2	强度五分类：`y_strength`，并输出各档概率（`predict_proba`）；同时保留回归派生的数值涨跌幅预测	Accuracy、macro-Precision/Recall/F1；概率分布输出
问题 3	阶段划分 + 阶段内关键因子组筛选，输出各阶段因子组权重（可解释）	阶段摘要表；Top-K 因子组及权重

符号约定与预测口径（避免信息泄露） 以月度为单位，记目标月价格为 y_t，特征为 $\mathbf{x}_t$。实际预测场景中，企业在 t 月决策时只能观测到截至 $t-1$ 月的信息（尤其进口、检修等存在公布滞后），因此本项目统一采用

$$\mathbf{x}_{t-1} \to y_t$$

即对所有特征做 `shift(1)`，用上一月信息预测下一月价格（预测步长 $t+1$）。后文所有 EDA 的“覆盖/可用性”统计亦以此建模口径为准。

2 数据说明与预处理

2.1 原始数据构成与业务维度

原始表大致可分为 5 类（详见表 2 的覆盖统计与第 3 节的文件溯源表）：

- **价格与预期：**现货价格（目标）、期货价格（快变量，覆盖较短）。
- **成本：**丙烯、PDH、乙烯裂解、MTO、CTO 等多条生产路线成本（高度共线但经济含义不同）。
- **供给：**PP 产量、进口量、检修损失、排产比例（结构性供给）。
- **需求：**塑编/注塑/BOPP 等下游开工率（景气与季节性明显）。
- **库存与宏观：**PP 石化库存（风险信号）、GDP（年度慢变量，作为背景景气）。

由于这些指标来自不同来源与口径，存在：频率不一致、缺失结构性强、部分变量公布存在滞后、且时间跨度差异显著。因此预处理的核心目标是：**把多源高频数据统一成“月度宽表”，并在严格滞后对齐下构造可用于预测的样本**。

2.2 覆盖差异与频率识别

表 2 给出各因子的时间覆盖、频率判断与覆盖月数（由 `scripts/run_pp_eda.py` 自动统计）。可以看到：核心长覆盖因子（价格、丙烯/乙烯成本、库存、主要开工率、产量、进口等）可支撑 2015-01∼2021-07 的长样本建模；而期货、PDH、检修、GDP 等变量覆盖较短，更适合作为“阶段/冲击”信号，需要配合缺失机制处理与敏感性对比。

表 2: 原始数据时间覆盖与频率概览（PP 数据/ 表 1–16；**口径说明：**“平均/最低/最高”表示原表包含平均、最低、最高三列；“单一数值列”表示原表每期仅提供一个数值字段，如库存或 GDP）

因子	口径	起始日期	结束日期	频率判断	覆盖 (月)
BOPP 开工率	平均/最低/最高	2010-01-01	2021-06-01	月度	138
CTO 成本	平均/最低/最高	2015-01-04	2021-07-15	日度/不规则	79
GDP	单一数值列	2016-12-31	2020-12-31	年度/稀疏	5
MTO 成本	平均/最低/最高	2016-01-04	2021-07-15	日度/不规则	67
PDH 成本	平均/最低/最高	2019-02-13	2021-07-15	日度/不规则	30
PP 产量	平均/最低/最高	2014-01-31	2021-06-30	月度	90
PP 价格	平均/最低/最高	2014-12-01	2021-07-15	日度/不规则	80
PP 开工率	平均/最低/最高	2014-01-03	2021-07-15	周度	91
PP 石化库存	单一数值列	2011-01-16	2021-07-15	日度/不规则	127
PP 进口量	平均/最低/最高	2010-01-31	2021-05-31	月度	137
丙烯成本	平均/最低/最高	2011-01-04	2021-07-15	日度/不规则	127
乙烯成本	平均/最低/最高	2014-12-31	2021-07-15	日度/不规则	80
塑编开工率	平均/最低/最高	2011-01-07	2021-07-15	周度	127
排产比例	平均/最低/最高	2014-01-02	2021-07-15	日度/不规则	91
期货价格	平均/最低/最高	2019-10-28	2021-07-15	日度/不规则	22
检修损失	平均/最低/最高	2018-01-31	2021-06-30	月度	42

2.3 统一月度化：周/日/年 → 月

本项目对每一张原始表统一生成月度特征口径（实现于 `pp_forecast/aggregation.py`）：

- 若表中包含 最低/最高/平均：以 平均作为主序列，月内计算 `mean/min/max/last`，并用 `range=max-min` 刻画月内波动。
- 若表中为单值序列（如库存、GDP）：同样计算 `mean/min/max/last/range`，以保持口径一致。

这样做的动机是：**月均值提供水平信息，而月内波动（range/last-mean）对拐点、冲击与风险状态更敏感**，可在后续特征工程中进一步利用。

2.4 时序对齐：$x(t-1) \to y(t)$ 与信息滞后

月度宽表生成后，我们将所有特征整体做 `shift(1)`，得到用于预测的输入 X。这一步非常关键：它把“能在 t 月末看到的数据”转换成“用于预测 t 月价格的可用信息”，从而与真实预测流程一致。

对低频宏观变量（GDP），我们先将年度值 **forward-fill** 到后续各月（直到下一次年度更新），再进行 `shift(1)`；这样 GDP 可作为“年度景气背景”特征进入月度模型，同时不引入未来信息。
对覆盖短且缺失结构明显的变量（期货、PDH、检修等），我们并行提供两种策略：默认在长样本下使用“中位数填补 + 缺失指示”保持样本量；在含期货方案中提供 `restrict` 模式，仅保留期货存在的月份做敏感性对比，避免将长段缺失完全依赖填补。

2.5 目标变量与标签构造（问题 1/2 共用）

以 **月均价口径**为默认（也支持月末价口径），构造：

- 价格回归目标：y_t（月度 PP 价格）
- 上月价格：y_{t-1}（记为 `y_prev`）
- 方向标签（严格口径）：若 $(y_t - y_{t-1}) > 0$ 为涨（1），否则为跌/不涨（0）
- 月度收益率：$r_t = (y_t - y_{t-1})/y_{t-1}$（记为 `y_return`）

其中 y_{t-1} 同时作为重要的滞后自回归项参与建模；方向标签用于问题 1 的“涨/跌”输出，也为问题 2 的强度分档提供基础。

2.6 涨跌强度分档（问题 2）

默认阈值(可配置):强波动阈值 5%(`strong_threshold=0.05`),平稳阈值 0.5%(`flat_threshold=0.005`)。据此将 r_t 分为 5 类：

`big_down`, `small_down`, `flat`, `small_up`, `big_up`

并在分类模型中输出各档概率（`predict_proba`）。

2.7 缺失处理、标准化与可复现流水线

缺失主要来源于：（1）覆盖短带来的结构性缺失；（2）高频数据缺测或节假日导致的自然缺失；（3）指标公布滞后。为在“样本量有限”的月度场景下获得稳健模型，本项目在 scikit-learn Pipeline 内对全部数值特征统一进行：

- 中位数填补（median imputation）+ 缺失指示特征
- 标准化（StandardScaler）

保证不同量纲特征可比，并显式利用“缺失机制”信息。该流水线对线性模型尤为重要（避免量纲支配系数），对树模型则主要用于保持一致的特征处理接口。

2.8 数据集版本与输出文件

为对比“期货信息”与“特征工程”的边际收益，我们构建四套数据集并在后续统一评测（默认输出到 `outputs/datasets/`）：

- `pp_base.csv`：不含期货，长样本基准；
- `pp_with_futures.csv`：含期货（覆盖短，支持 `restrict`）；
- `pp_base_engineered.csv`：在基准上加入派生特征工程；
- `pp_with_futures_engineered.csv`：含期货 + 特征工程。

数据集构建命令见附录 A，并在问题 1/2 中对四套方案进行并行评估。

3 特征工程与描述性统计（时间序列 EDA）

本节对应题目对“特征工程与描述性统计分析”的要求。我们同时对 **原始数据**、**月度聚合后的宽表**、以及 **建模数据集**做统计与可视化（输出均为 PDF）。

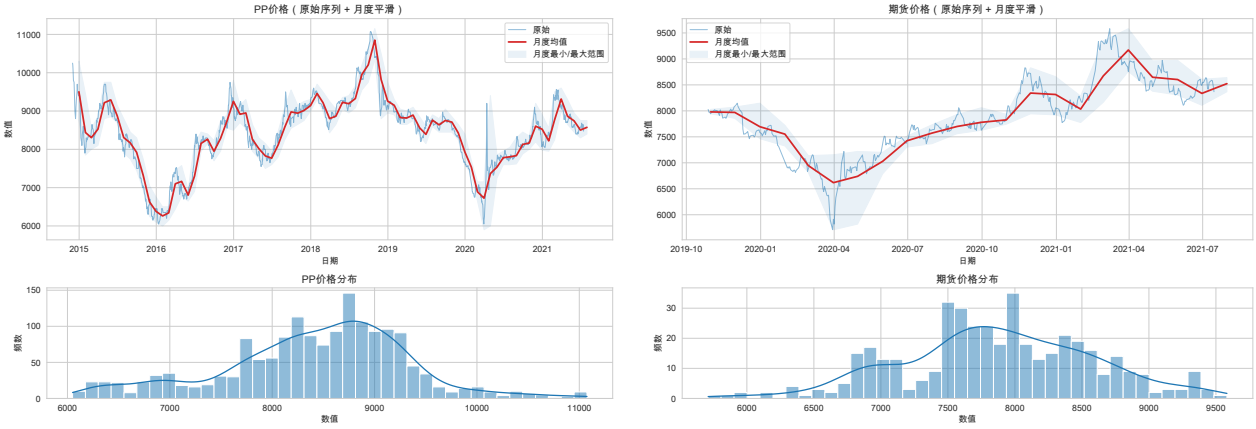
3.1 原始数据概览

我们对 PPData 的每张原始表都生成了“原始序列 + 月度平滑 + 分布直方图”(路径:outputs/eda/raw/figu 用于检查频率差异、异常值与分布形态。为避免“图在上一页、解释在下一页”带来的空白，本节采用“**一行两图 + 紧接文字解读**”的排版方式：先展示同类因子（两张图），再分别解释每个因子的序列特征与业务含义。

表 3 汇总了本节每段对应的原始数据文件路径，便于溯源与复现。

表 3: 原始数据段落与文件路径对应关系（PP 数据/ 表 0–16）

段落/因子	数据文件路径
汇总表（简要信息）	PP 数据/0-pp 价格预测简要信息汇总（案例）1011.xlsx
PP 价格（现货）	PP 数据/1-华东市场 PP 粒市场价_ 法定工作日.xlsx
PP 产量	PP 数据/2-中国 PP 月度产量.xlsx
塑编开工率	PP 数据/3-塑编行业周度开工率.xlsx
排产比例（拉丝级生产比例）	PP 数据/4-PP 拉丝级生产比例日度数据.xlsx
PP 进口量	PP 数据/5-PP 进口量月度数据_ 滞后一月更新.xlsx
PP 开工率（注塑制品）	PP 数据/6-PP 注塑制品周度开工率.xlsx
BOPP 开工率	PP 数据/7-BOPP 月度开工率.xlsx
PDH 成本	PP 数据/8-PDH 生产路线日度含税成本.xlsx
乙烯成本（乙烯裂解路线）	PP 数据/9-乙烯裂解生产路线日度含税成本.xlsx
MTO 成本	PP 数据/10-MTO 生产路线日度含税成本.xlsx
CTO 成本	PP 数据/11-CTO 生产路线日度含税成本.xlsx
丙烯成本（外采丙烯路线）	PP 数据/12-外采丙烯生产路线日度含税成本.xlsx
期货价格（大商所 PP）	PP 数据/13-大商所 PP 期货价格_ 短数据.xlsx
检修损失	PP 数据/14-PP 月度检修实际损失量.xlsx
PP 石化库存	PP 数据/15-pp 石化库存_ 每 3 个工作日更新.xlsx
GDP（年度）	PP 数据/16-年度 GDP.xlsx



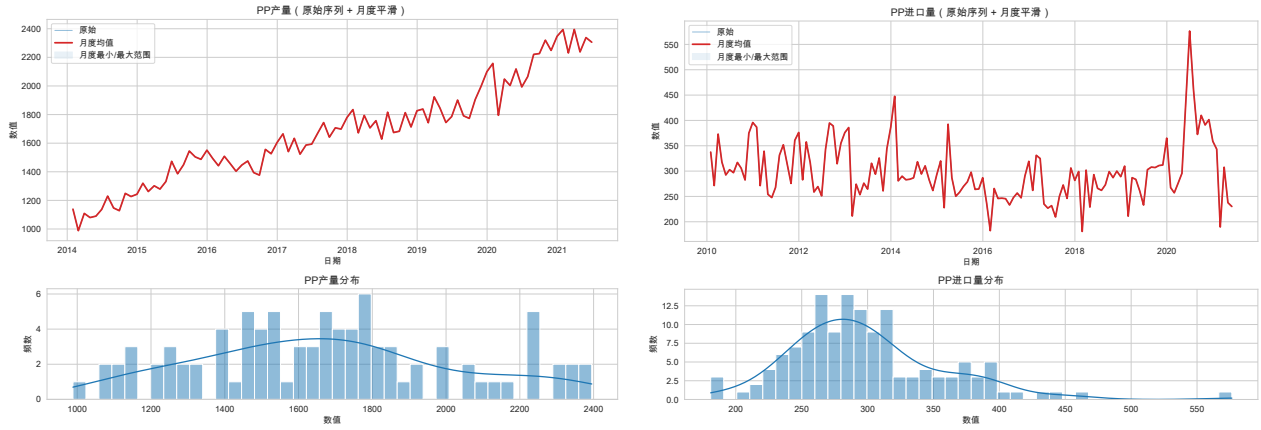
(a) PP 价格（法定工作日） (b) 期货价格（短数据）

图 1: 原始数据：价格相关因子

PP 价格（现货）：覆盖 2014-12 至 2021-07（约 80 个月），日度/不规则数据；均值约 8413、标准差约 907，取值范围约为 [6050, 11075]。时序上呈现明显的周期波动与波动聚集：2015–2016 年低位下行，随后进入多轮上行与回落，2018 年末出现阶段性高位，2020 年初出现急跌后修复。图中浅色

“月度最小/最大范围”在拐点与冲击月份显著变宽，提示月内波动状态（range、max-min）本身携带有效信息；分布上主峰集中在中高价区间且右尾更长，说明极端高价月份虽少但会显著抬升均值并加大波动，从而提高均值预测难度。

期货价格：覆盖 2019-10 至 2021-07（约 22 个月），均值约 7866、标准差约 707，范围约为 [5711, 9582]，且存在少量缺失（约 1.4%）。时序上与现货同向且对拐点更敏感（2020 年急跌、2021 年快速上行阶段更为明显）；在月度聚合层面，期货与现货价格的斯皮尔曼相关在 lag0 约 0.97、lag1 仍约 0.89，表明其对价格水平解释力很强。业务上期货更接近“预期/风险偏好”的快变量，但覆盖期短会缩短有效样本，因此我们保留“纳入/不纳入期货”两套方案并进行对比评估。

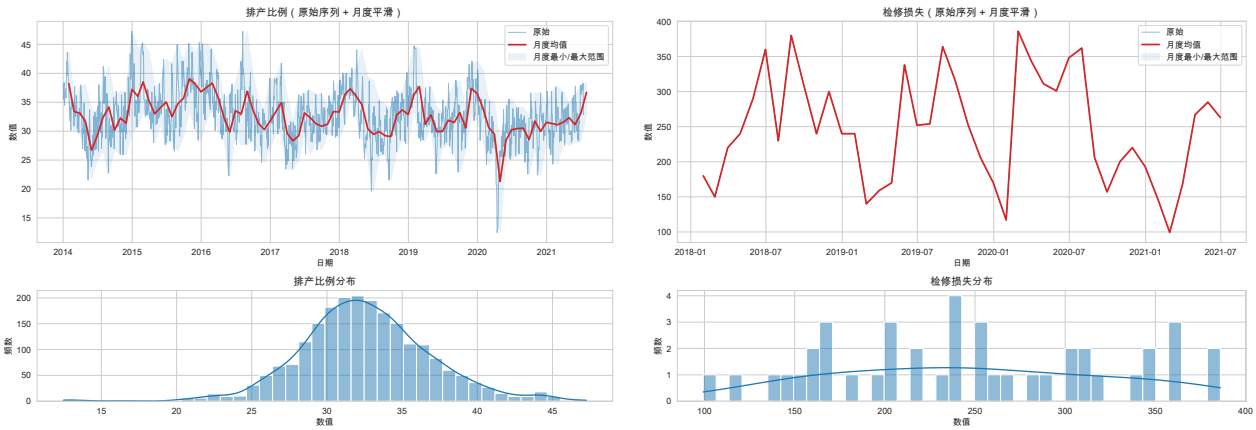


(a) PP 产量（月度）　　(b) PP 进口量（月度，滞后一月更新）

图 2: 原始数据：供需相关因子

PP 产量：覆盖 2014-01 至 2021-06（90 个月），月度统计；均值约 1676、标准差约 348，范围约为 [988, 2395]。序列整体趋势上行，反映国内产能投放与装置稳定运行能力提升；月度波动并非随机噪声，往往对应检修、装置开停与需求旺淡季。由于产量具有趋势项，其与价格的简单相关并不强（例如月度聚合后 lag0 相关约 0.10、lag1 约 0.09），说明“供给水平”对价格的影响往往需要与成本、库存、需求共同解释，单独使用可能产生误判。

PP 进口量：覆盖 2010-01 至 2021-05（137 个月），月度统计且源数据标注“滞后一月更新”；均值约 299、标准差约 59，范围约为 [181, 576]。分布明显右偏并存在尖峰（少数月份进口量显著高于常态），体现外盘资源、国际价差与到港节奏的集中性。月度聚合后其与价格在较长滞后处出现负相关（例如 lag6 约 -0.13），符合“高价吸引进口、到港存在运输滞后”的机制。建模上更适合使用相对量与变化率（如 供需 __ 进口 _div_ 产量 __mean、动量/滚动均值），并严格用上一月信息预测下一月以规避更新滞后带来的信息穿越。

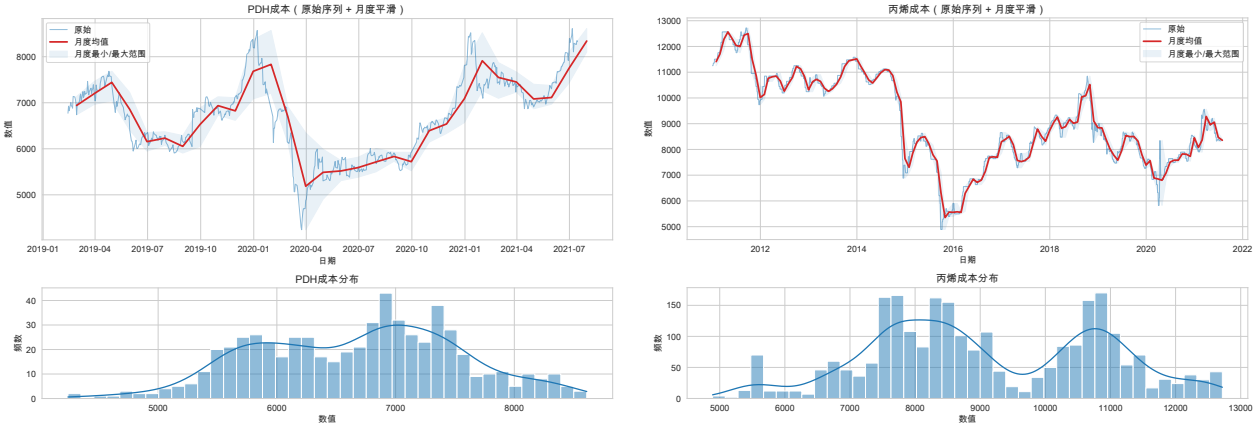


(a) 排产比例（日度） (b) 检修损失（月度）

图 3: 原始数据：供给结构与检修因子

排产比例（拉丝级生产比例）： 覆盖 2014-01 至 2021-07（约 91 个月），日度/不规则数据；均值约 32.5、标准差约 4.3，范围约为 [12.5, 47.3]。该指标刻画 “PP 产出结构” 而非总量：在总产量相近时，拉丝级比例上升往往意味着下游（塑编等）需求结构更偏向拉丝品种。其与价格的简单相关整体偏弱且在短滞后处为负（例如 lag0 约 -0.02、lag1 约 -0.16），提示排产比例可能更多是对价格/利润与订单结构的被动调整而非单向驱动；因此更适合以 “变化率/偏离度”（例如环比变化、滚动 z-score）以及与下游开工的联动特征进入模型。

检修损失： 覆盖 2018-01 至 2021-06（42 个月），月度统计；均值约 247、标准差约 76.5，范围约为 [99, 386]。从业务逻辑看，检修损失代表供给收缩，应对价格形成支撑；但月度聚合后其与价格在短滞后处相关并不显著且略偏负（例如 lag0/lag1 约 -0.09），可能反映 “低景气期更集中检修” 的反向因果以及样本期较短（仅 42 个月）带来的不稳定。建模上更适合将其视为阶段性供给冲击信号，并配合成本、库存共同解释。



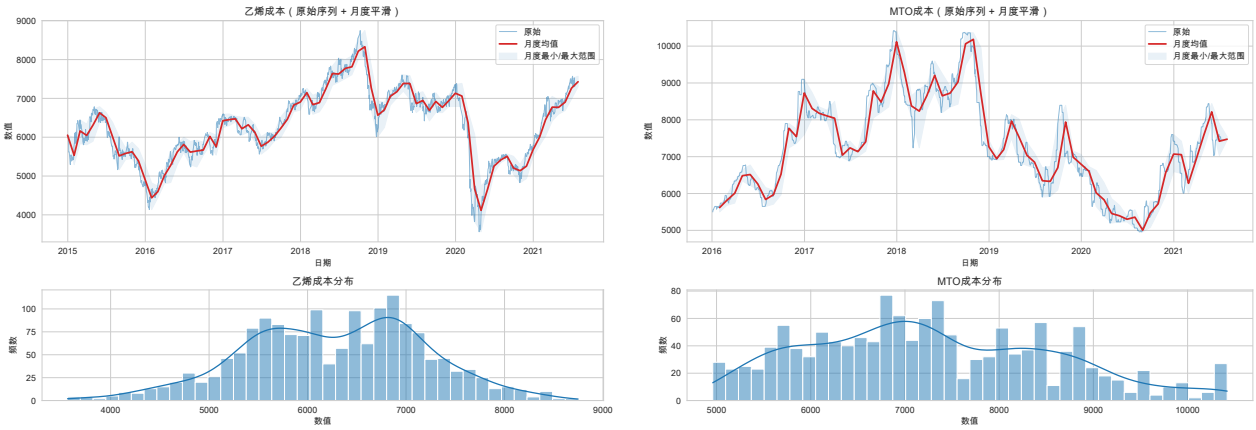
(a) PDH 成本（日度） (b) 丙烯成本（日度）

图 4: 原始数据：成本相关因子

PDH 成本： 覆盖 2019-02 至 2021-07（约 30 个月），日度数据；均值约 6677、标准差约 848，范围约为 [4240, 8610]。2020 年前后存在明显的断崖式下跌与随后修复，体现能源/原料端冲击的快速传导。月度聚合后其与价格在短滞后处呈中等正相关（lag0 约 0.46），但由于覆盖期短，相关结构更容易受少数极端月份影响，需通过正则化、集成与稳健特征（价差/比值）控制不确定性。

丙烯成本： 覆盖 2011-01 至 2021-07（约 127 个月），日度数据；均值约 9082、标准差约 1760，范围约为 [4891, 12718]。时序中出现多次 “制度切换”（不同宏观/能源状态下的成本区间），分布呈多峰

与厚尾；与 PP 价格的相关性极强（月度聚合后 lag0 约 0.89、lag1 仍约 0.85），符合“成本驱动定价”的产业逻辑。建模上应重点刻画利润空间（现货-成本价差）与成本冲击的持续性（滚动均值/波动/动量），以提升对成本行情的识别。

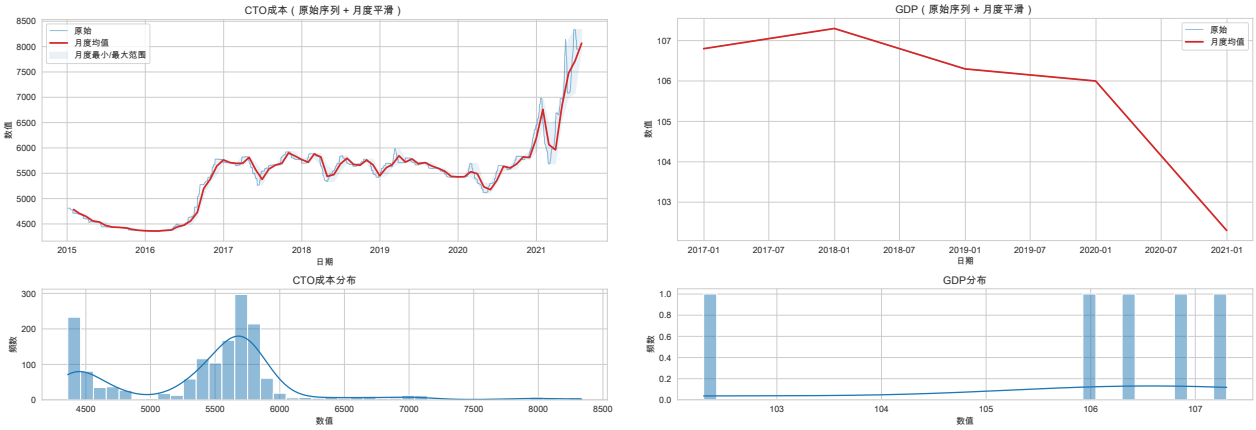


(a) 乙烯成本（日度） (b) MTO 成本（日度）

图 5: 原始数据：其他成本路线因子

乙烯成本：覆盖 2014-12 至 2021-07（约 80 个月），日度/不规则数据；均值约 6276、标准差约 913，范围约为 [3563, 8747]。该指标代表乙烯裂解路线成本，与能源价格与化工景气高度相关；月度聚合后其与 PP 价格呈显著正相关（lag0 约 0.73、lag1 约 0.65），体现替代品/上游共同驱动的联动行情。考虑到成本类因子之间高度共线，建模中应配合正则化或使用价差/比值特征降低冗余。

MTO 成本：覆盖 2016-01 至 2021-07（约 67 个月），日度/不规则数据；均值约 7277、标准差约 1293，范围约为 [4960, 10420]。序列波动幅度大且与价格上行周期高度同步；月度聚合后与 PP 价格的相关性较强（lag0 约 0.79、lag1 约 0.74）。由于覆盖期相对较短且波动受煤价/甲醇价与政策扰动影响明显，建议与其他成本路线共同构造“成本上行压力指数”或利润特征进入模型。



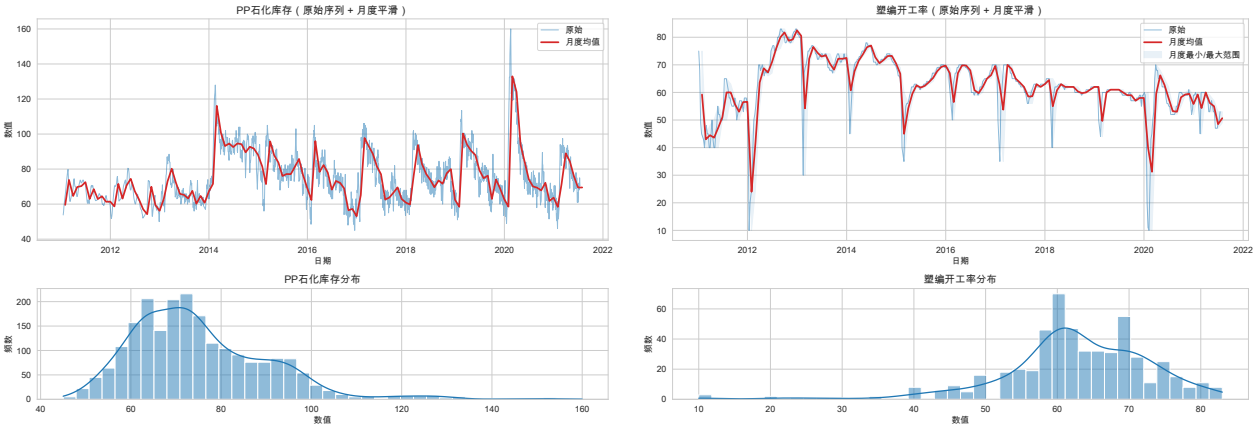
(a) CTO 成本（日度） (b) GDP（年度）

图 6: 原始数据：煤化工与宏观因子

CTO 成本：覆盖 2015-01 至 2021-07（约 79 个月），日度/不规则数据；均值约 5436、标准差约 703，范围约为 [4360, 8335]。与 PP 价格同样呈周期波动，但整体相关强度弱于丙烯/MTO（例如 lag0/lag1 约 0.55）；这可能与 CTO 路线在成本结构与行业边际供给中的地位有关。作为补充成本因子，其更适合与其他路线共同使用，以刻画“不同工艺的边际成本与供给弹性”。

GDP：覆盖 2016–2020（5 个年度观测），年度数据；均值约 105.74、标准差约 1.78，范围约为 [102.3, 107.3]。考虑到 GDP 频率较低但可作为宏观“慢变量”，我们在构建月度宽表时将年度值

forward-fill 到后续各月（直到下一次年度更新），并仍保持 $\mathbf{x}_{t-1} \to y_t$ 的滞后对齐以避免信息泄露。这样 GDP 自 2017-01 起可作为“年度景气水平”背景特征进入月度模型，但其阶梯式变化决定了对短期波动的解释力有限，更适合与趋势/阶段变量共同使用，或通过差分刻画年度切换带来的边际影响。

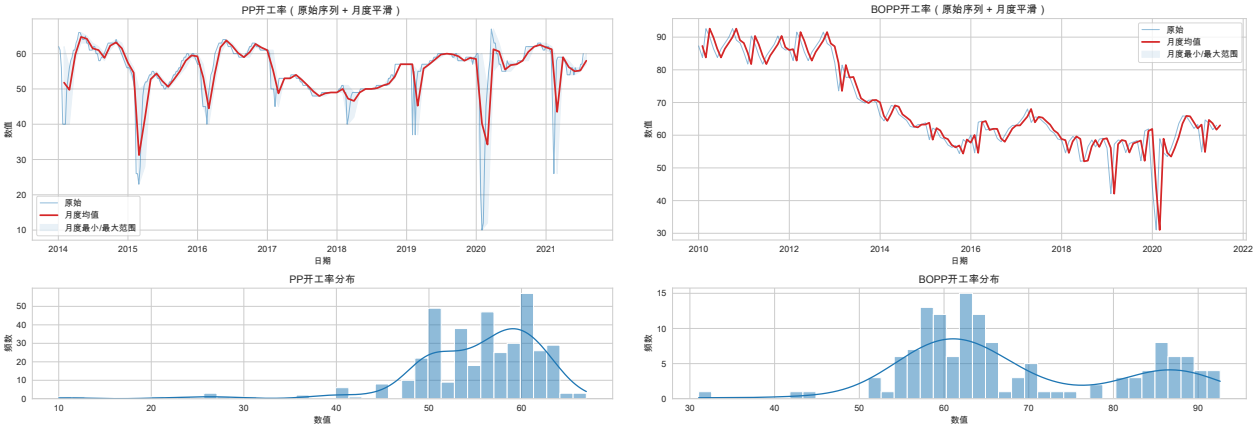


(a) PP 石化库存（每 3 个工作日更新） (b) 塑编开工率（周度）

图 7: 原始数据：库存与下游景气因子

PP 石化库存：覆盖 2011-01 至 2021-07（约 127 个月），高频更新（每 3 个工作日）；均值约 75.1、标准差约 15.2，范围约为 [45, 160]。库存序列短期起伏明显且右尾较长，存在少数尖峰（累库冲击），说明库存更像“风险信号”：当库存快速上冲时，价格往往承压或进入震荡。月度聚合后其与价格在短滞后呈负相关（lag1 约 -0.17、lag2 约 -0.22），方向符合经验，但强度有限，提示其效应可能依赖成本与需求状态（例如高成本下库存上升对价格的影响更大）。因此我们在特征工程中使用库存的 `last/range` 与比值特征来增强信号稳定性。

塑编开工率：覆盖 2011-01 至 2021-07（约 127 个月），周度数据；均值约 62.6、标准差约 10.5，范围约为 [10, 83]。序列具有显著季节性（年初/春节附近低谷）并伴随少量异常低值点；长期中枢存在缓慢下移迹象，可能反映行业景气与产能利用率变化。其与价格的短滞后相关偏弱（lag0 约 -0.11），但在更长滞后处相关有所增强（例如 lag8 约 0.18），提示下游景气对价格的影响可能存在传导滞后。建模上更适合通过滚动统计、动量与“下游开工指数”类聚合特征来表征需求持续性，而非仅用单月水平值。



(a) PP 开工率（注塑制品，周度） (b) BOPP 开工率（月度）

图 8: 原始数据：其他下游景气因子

PP 开工率（注塑制品）：覆盖 2014-01 至 2021-07（约 91 个月），周度数据；均值约 55.1、标准差

约 7.5，范围约为 [10, 67]。序列存在明显的节假日季节性（低谷月份集中）与阶段性回落；其与价格在短滞后处呈负相关（例如 lag0 约 -0.29），更像“价格反向压制下游开工”的表现而非单向需求驱动。为提升预测价值，建模中应优先使用上一月/多月平均与变化率（动量、滚动均值差）来刻画需求景气的持续性。

BOPP 开工率：覆盖 2010-01 至 2021-06（138 个月），月度统计；均值约 68.3、标准差约 12.9，范围约为 [31.0, 92.6]。长期看呈现明显周期与季节波动，但与 PP 价格的同步相关较弱（lag0 约 -0.06）；在中等滞后下相关上升（例如 lag6–lag8 约 0.29–0.33），提示其可能刻画更偏“终端需求/景气周期”的慢变量，对价格的影响存在传导与库存缓冲。该特征更适合作为阶段判断或与其他下游开工共同构造综合指数使用。

3.2 月度特征覆盖与缺失

数据口径与“覆盖”的定义 由于原始数据同时包含日度/周度/月度/年度序列，我们先将每张表统一聚合为**月度特征**（mean/min/max/last/range 五种口径）。在建模时统一采用 $\mathbf{x}_{t-1} \to y_t$ 以避免信息泄露，因此本小节的“可用性/覆盖”统一按**建模口径**统计：对每个预测月份 t，若上一月 $t-1$ 的该因子月度均值存在，则该因子在 t 月可用（记为 1），否则记为缺失（0）。

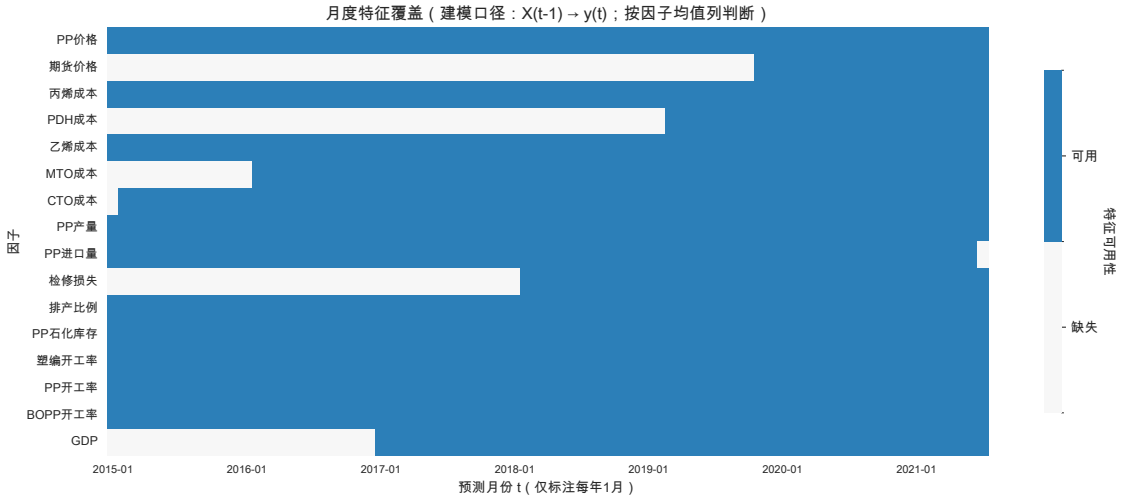
建模窗口 由于目标价格（PP 价格）最早可用于预测的月份为 2015-01（2014-12 用作 $t-1$ 特征），本项目的月度建模窗口为 2015-01 至 2021-07（共 79 个月）。表 4 汇总各因子在该窗口内的首次/最后可用月份、可用月数与缺失率。

表 4: 建模窗口的月度特征可用性汇总（预测月 t 使用 $t-1$ 月特征；共 79 个月）

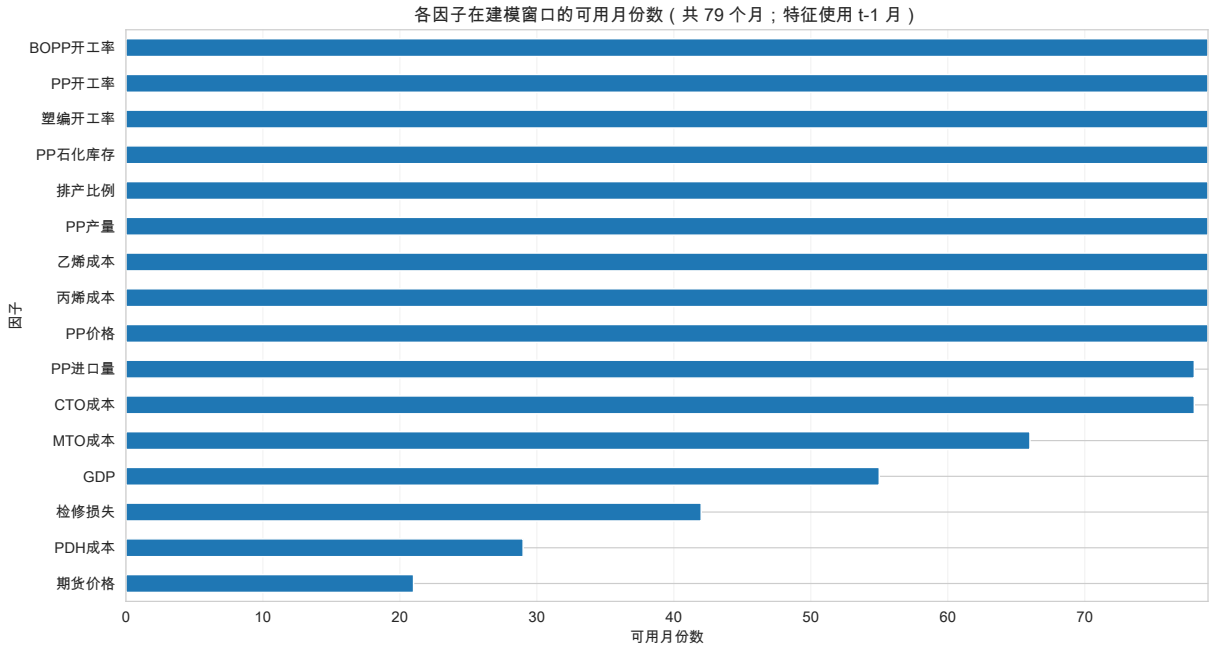
因子	首次可用月	最后可用月	可用 (月)	缺失 (月)	缺失率
PP 价格	2015-01	2021-07	79	0	0.0%
期货价格	2019-11	2021-07	21	58	73.4%
丙烯成本	2015-01	2021-07	79	0	0.0%
PDH 成本	2019-03	2021-07	29	50	63.3%
乙烯成本	2015-01	2021-07	79	0	0.0%
MTO 成本	2016-02	2021-07	66	13	16.5%
CTO 成本	2015-02	2021-07	78	1	1.3%
PP 产量	2015-01	2021-07	79	0	0.0%
PP 进口量	2015-01	2021-06	78	1	1.3%
检修损失	2018-02	2021-07	42	37	46.8%
排产比例	2015-01	2021-07	79	0	0.0%
PP 石化库存	2015-01	2021-07	79	0	0.0%
塑编开工率	2015-01	2021-07	79	0	0.0%
PP 开工率	2015-01	2021-07	79	0	0.0%
BOPP 开工率	2015-01	2021-07	79	0	0.0%
GDP	2017-01	2021-07	55	24	30.4%

如何读图 图 9 上半部分为“特征可用性热力图”：**行 = 因子，列 = 预测月份** t，颜色表示该月是否存在 x_{t-1}（上一月信息）。因此，像“期货价格/PDH 成本/检修损失”这类覆盖较短的因子，会在时间轴上呈现“从某一时点开始出现的色块”；而 GDP 虽为年度数据，但我们将其 forward-fill 到后续月份，所以从 2017-01 起呈现连续可用（2015-01–2016-12 仍缺失）。

关键结论（对建模的影响） （1）多数核心因子（现货价格、丙烯/乙烯成本、库存、主要开工率、产量等）在 2015-01–2021-07 的窗口内几乎**全覆盖**，因此长样本建模的主要信息来源稳定；（2）少数因子存在**结构性缺失**：例如期货价格仅自 2019-11 起可用（受 $t-1$ 滞后影响），PDH 成本自 2019-03 起可用，检修损失自 2018-02 起可用；GDP 作为年度数据虽已 forward-fill 到后续月份，但在 2017 年之前仍缺失。对这些短覆盖特征，本项目采取两种策略并行：一是通过“中位数填补 + 缺失指示”在**长样本**下使用它们（缺失本身也可能携带阶段信息），二是在“含期货”方案中提供 `restrict` 模式，仅保留期货存在的月份做敏感性对比，避免将大段缺失完全依赖填补。



(a) 特征可用性热力图（建模口径：x_{t-1} 是否存在）



(b) 各因子可用月份数（建模窗口内统计）

图 9: 月度特征覆盖与缺失（由 `scripts/run_pp_eda.py` 生成；建模口径：$\mathbf{x}_{t-1} \to y_t$）

3.3 平稳性检验与变换建议

表 5 对月度因子均值序列做 ADF 单位根检验，并同时报告“一阶差分”后的 p 值，用于判断是否存在明显趋势项/随机游走成分。多数价格与成本类因子在水平值上难以拒绝单位根假设（$p > 0.05$），

但差分后显著平稳（$p \ll 0.05$）；这意味着在建模中，**变化信息（动量/收益率/价差的变化）往往比绝对水平更稳定**。因此，我们在可选的特征工程开关中为每个因子构造了 `pct_change` 动量与滚动统计，并对成本路线进一步构造“现货-成本”的价差/比值，以增强对拐点与阶段切换的刻画能力。

表 5: 月度因子平稳性检验（ADF；水平值 vs 一阶差分）

因子	ADF p（水平）	ADF p（差分）	样本数（月）	差分样本数（月）
PP 价格	0.080	<0.001	80	79
期货价格	0.517	0.023	22	21
丙烯成本	0.204	<0.001	127	126
PDH 成本	0.255	0.004	30	29
乙烯成本	0.219	<0.001	80	79
MTO 成本	0.112	<0.001	67	66
CTO 成本	0.988	0.012	79	78
PP 产量	0.892	<0.001	90	89
PP 进口量	<0.001	<0.001	137	136
检修损失	0.002	<0.001	42	41
排产比例	<0.001	<0.001	91	90
PP 石化库存	<0.001	0.004	127	126
塑编开工率	0.010	<0.001	127	126
PP 开工率	0.060	0.021	91	90
BOPP 开工率	0.425	0.025	138	137
GDP	0.947	<0.001	56	55

备注： 期货由于覆盖较短，ADF 与相关分析更容易受少数月份影响；GDP 虽 forward-fill 后具备更长的月度序列，但其“阶梯型”变化使水平值通常仍表现为非平稳（需差分后更接近平稳）。因此这类慢变量更适合用作背景信号/阶段解释或与趋势特征结合，而不宜单独与高频因子做强因果解读。

3.4 因子共线性与信息冗余

图 10 展示了各因子月度均值的 Spearman 相关矩阵，用于识别“信息冗余”（高度共线）与潜在的替代关系。可以看到，多条成本路线之间存在显著正相关（如丙烯/乙烯/MTO/CTO），且与 PP 价格本身也高度同向；这符合产业逻辑，但也意味着把多条成本水平**同时**作为输入会引入较强共线性：线性模型中会导致系数不稳定、方差膨胀，树模型中则可能出现“重要性被分摊”。因此，本项目在建模上同时使用了（1）Ridge/Huber 等正则化/稳健线性模型，（2）树模型与集成学习，以及（3）更具经济含义的价差/比值特征，来降低共线性的负面影响并提升可解释性。

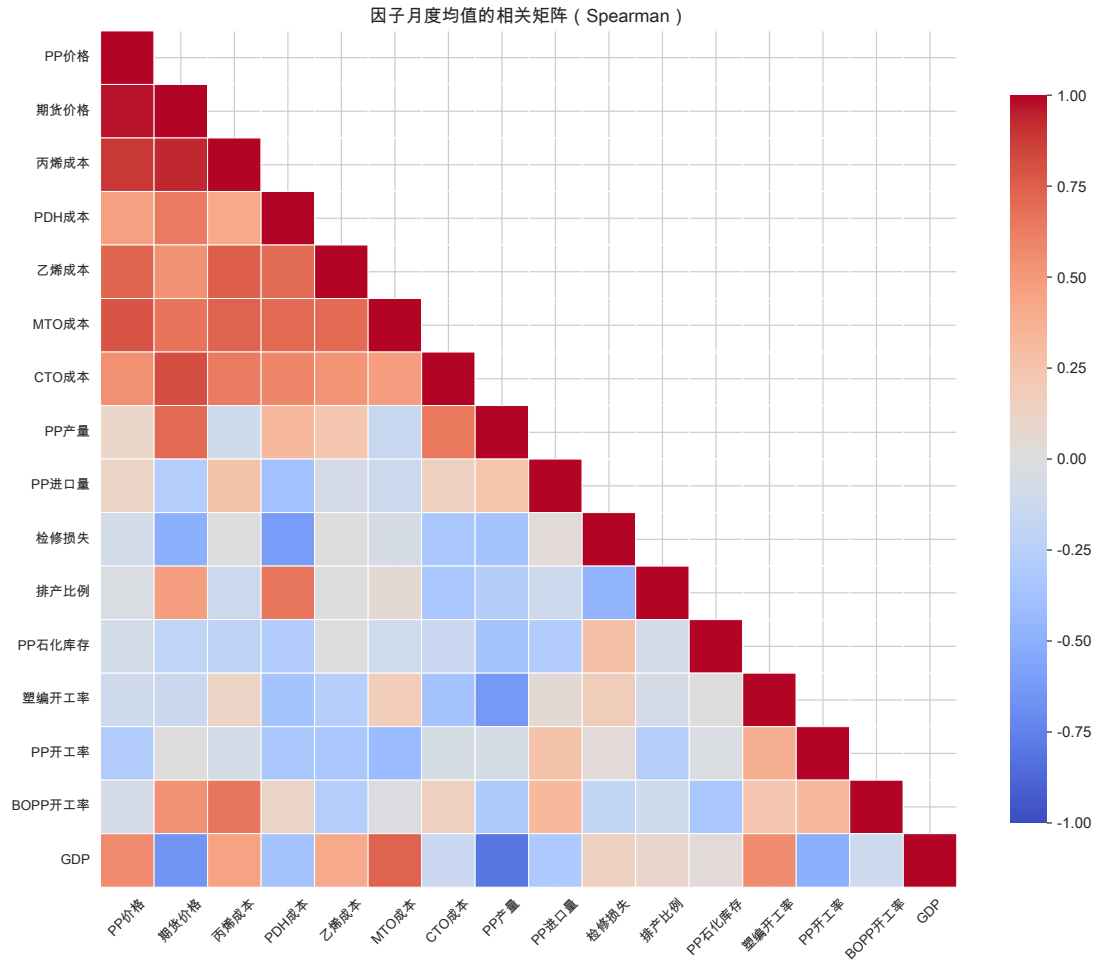


图 10: 因子月度均值相关矩阵（Spearman；下三角显示）

3.5 关系稳定性：滚动滞后相关

即便总体相关较高，因子与价格关系也可能随阶段发生变化（例如库存与开工率常受“价格反向压制需求/供给调整”影响）。为检查这种不稳定性，我们对若干关键因子计算了滚动窗口的滞后相关（以 `shift(1)` 对齐：用 x_{t-1} 与 y_t 相关，避免信息泄露），结果如图 11。总体上，成本类因子对价格的滞后相关相对稳定且大多为正，但强度会随行情周期起伏；而库存、产量、开工率等“量”变量的相关方向与强度更容易出现阶段性翻转，提示其效应依赖于成本、库存与需求状态的组合。这也解释了为什么在问题 3 中需要进行**分阶段因子筛选**：与其假设“全样本一套固定权重”，不如在不同阶段分别选择更有效的因子组，并通过正则化系数/集成权重来获得更稳健的影响度量。

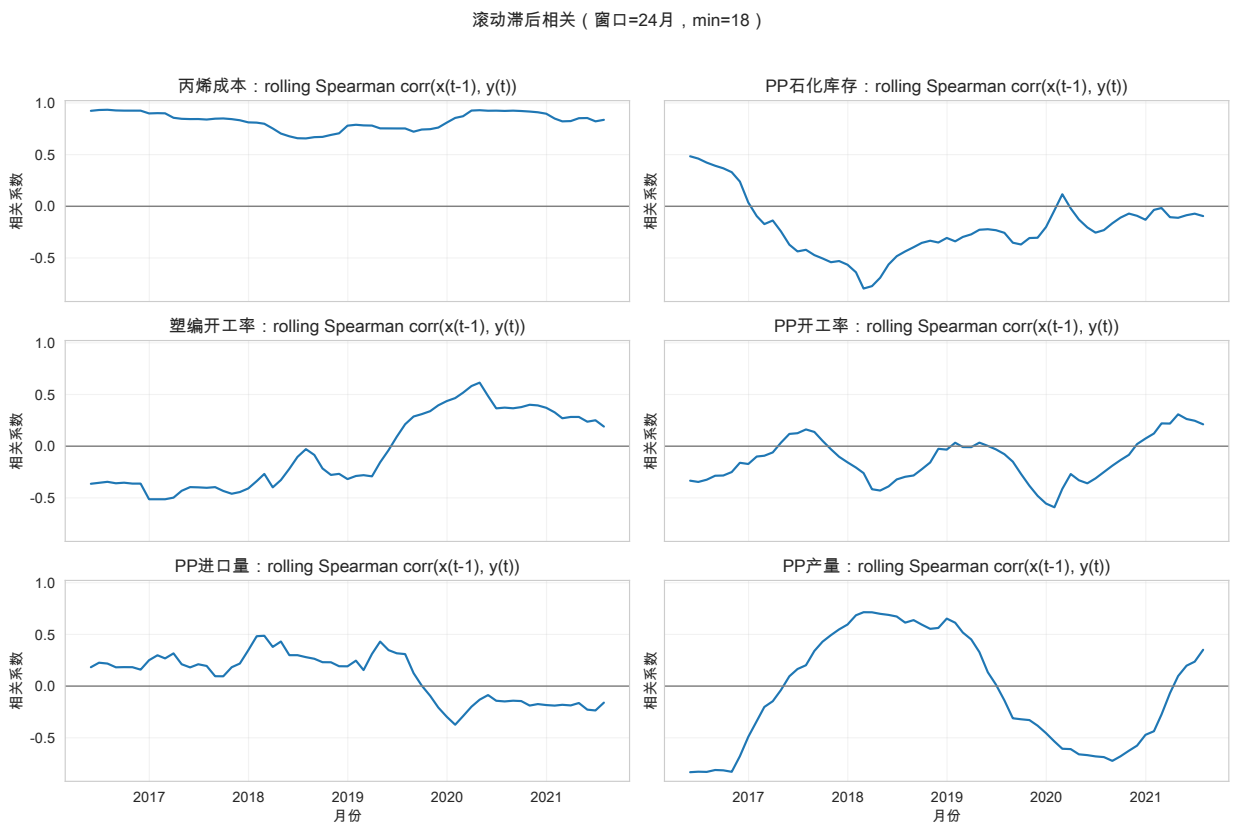


图 11: 滚动滞后相关（rolling Spearman corr$(x(t-1), y(t))$；窗口 =24 月）

3.6 建模数据集（pp_base）EDA

样本窗口与目标分布 `pp_base` 的建模窗口为 2015-01 至 2021-07（79 个月）。目标价格 y 的均值约 8398、标准差约 881；月度收益率 y_return 的标准差约 4.8%，极值约为 [-11.2%, 12.0%]，呈现一定厚尾与“波动聚集”，意味着单纯的均值回归在行情冲击月份更易失真。方向标签（涨/跌）在全样本上几乎均衡（40 vs 39），但强度五分类存在不均衡（`flat` 较少），因此问题 2 更适合使用 macro-F1 以及概率输出进行评估。

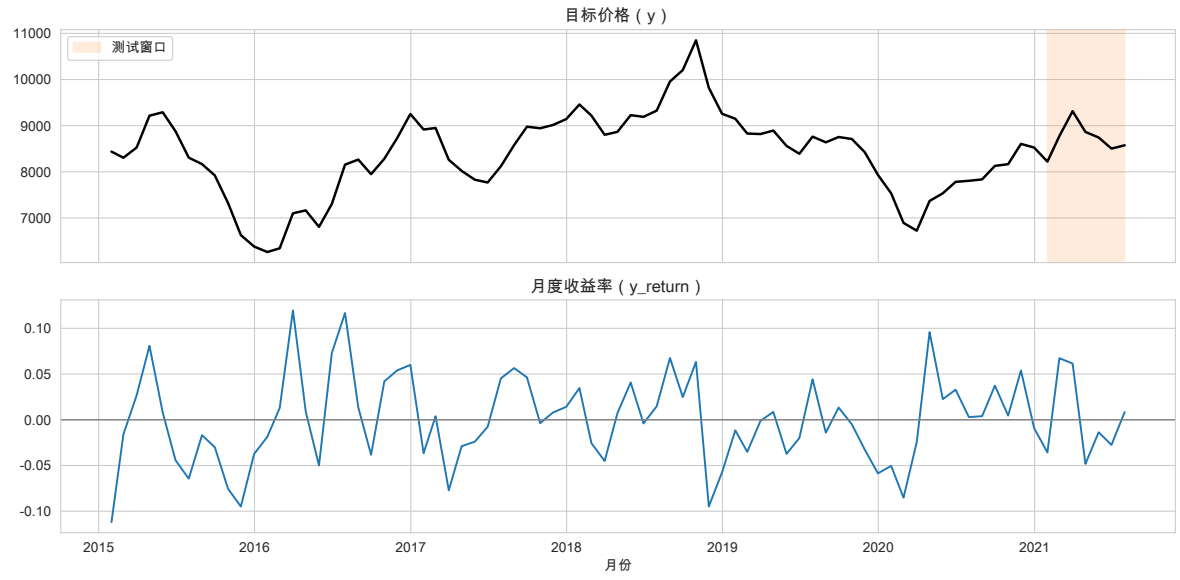
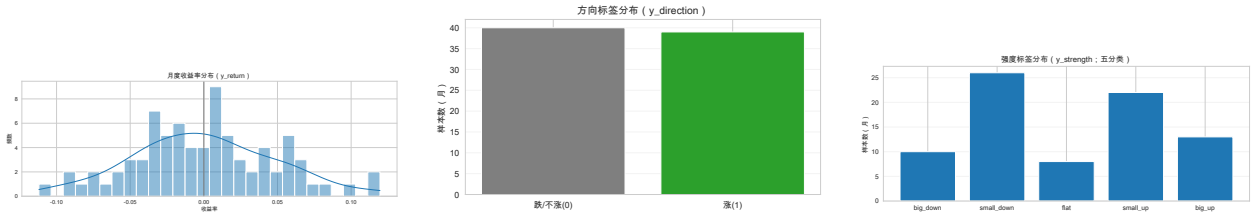


图 12: 目标序列：y 与 y_return（含测试窗阴影）

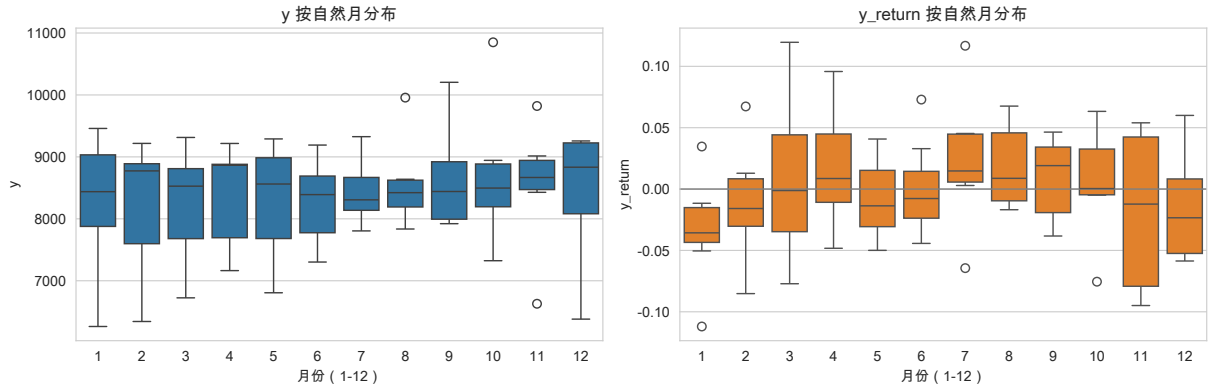


(a) y_return 分布 (b) 方向标签 (c) 强度五分类

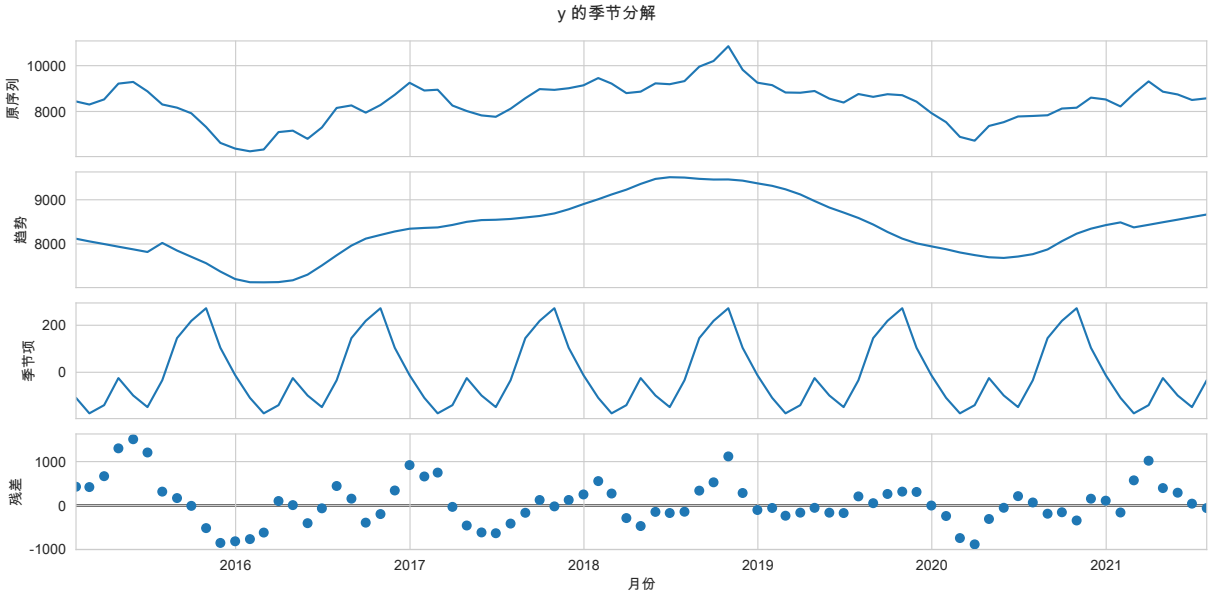
图 13: 目标与标签分布（pp_base）

季节性、趋势与自相关 图 14 的箱线图显示价格存在一定“月份效应”，而收益率的季节性更弱，提示用日历特征（月份/正余弦）去刻画长期规律更合适。季节分解进一步将序列拆解为趋势与季节项：趋势项平滑、季节项相对稳定，说明月度价格同时受“长期景气/成本中枢”与“周期性需求/库存变化”共同驱动。

另一方面，y 在 ACF 上通常表现为强自相关，而 y_return 的自相关显著更弱（图 15），这也解释了本项目在建模上采用两条思路并行：（1）对 y 做点预测并从中派生方向/强度；（2）直接对 y_return 的分档标签建模输出概率。

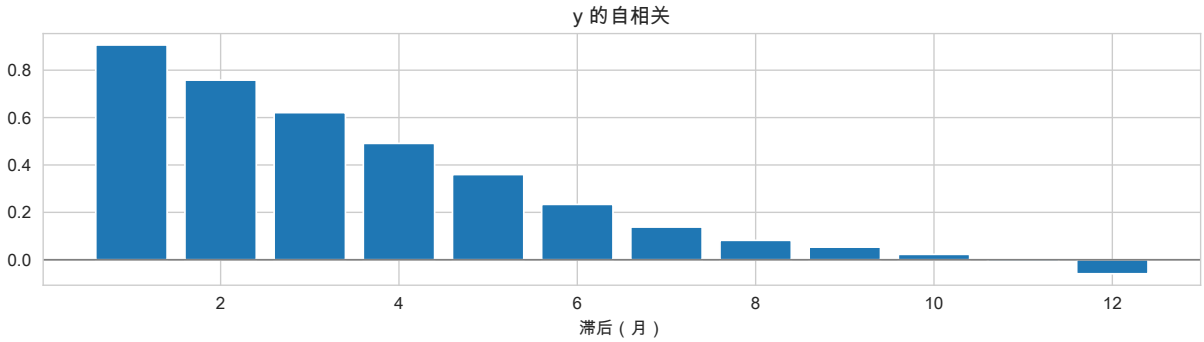


(a) 按月份的季节性箱线图

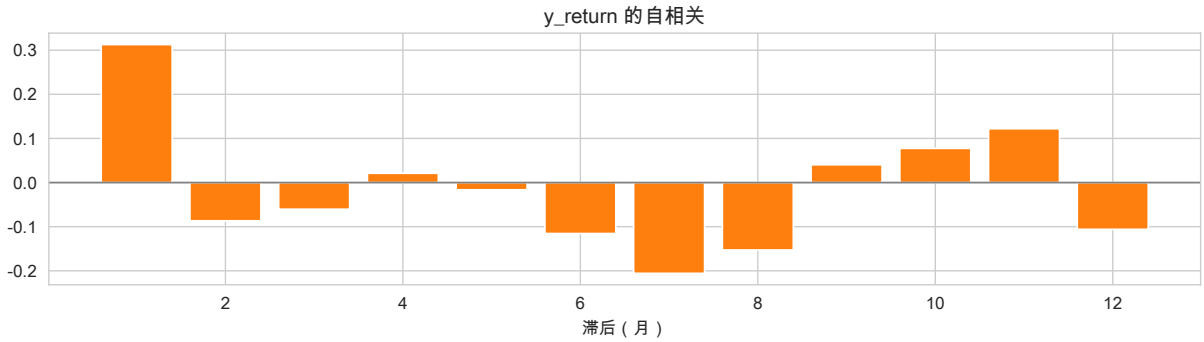


(b) y 的季节分解（趋势/季节/残差）

图 14: 季节性与趋势分解（pp_base）



(a) 价格水平 y 的自相关（ACF）



(b) 收益率 y_return 的自相关（ACF）

图 15: 目标序列的自相关结构（pp_base）

缺失结构与“结构性缺失” 图 16 从因子组角度给出了逐月缺失率。可以看到，多数因子在 2015-01～2021-07 近乎连续覆盖；而短覆盖因子（期货、PDH、检修、GDP 等）会形成“整段缺失”，这类缺失并非随机噪声，更像“该阶段不可观测/未采集”的结构性信息。为保证长样本可用性，本项目默认采用“中位数填补 + 缺失指示”将缺失机制显式纳入模型；同时提供 `restrict` 方案仅保留特征存在的月份用于敏感性对比。

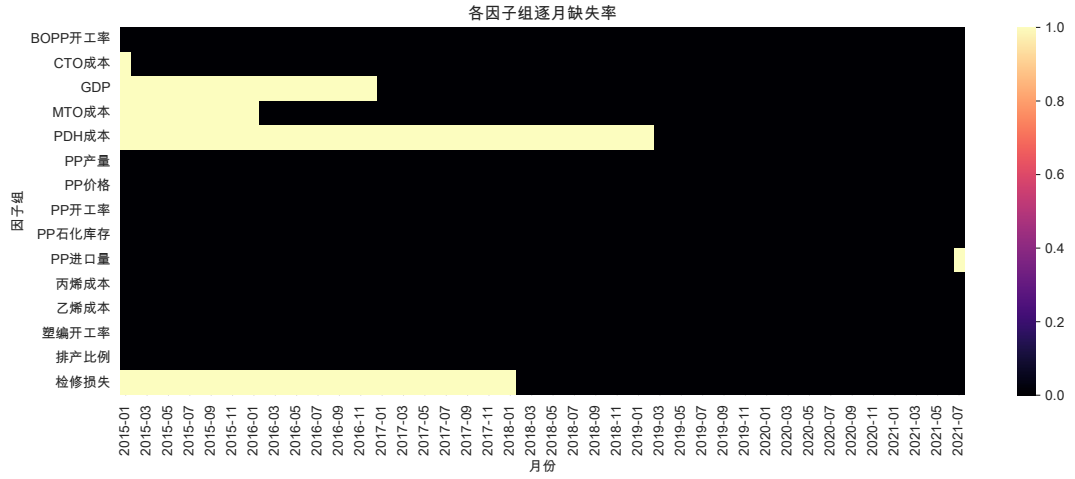
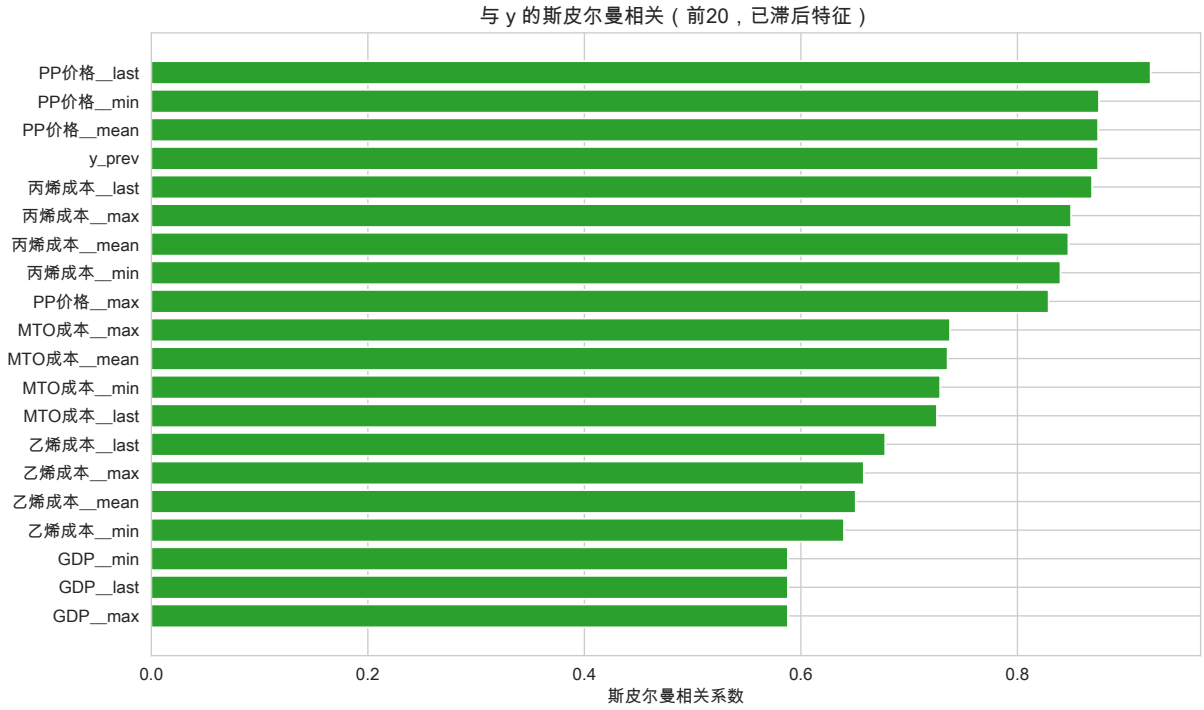


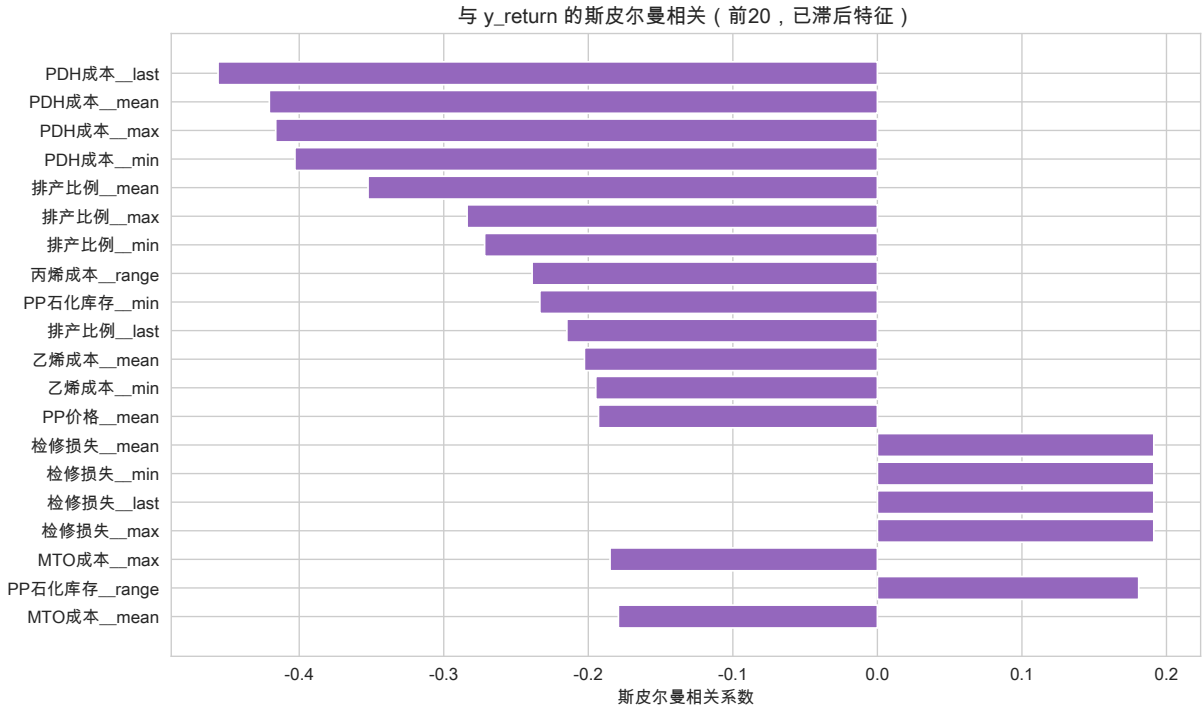
图 16: 按因子组统计的逐月缺失率热力图（pp_base）

与目标的相关性：水平值 vs 变化值 图 17 对比了“与 y”以及“与 y_return”的相关性 Top-20。可以看到：（1）水平值 y 的相关性主要由成本类（尤其丙烯）与价格自身的滞后项解释，反映“成

本驱动 + 价格惯性”的主逻辑；(2) 收益率 y_return 的相关性整体更弱、且更依赖短覆盖的冲击因子与结构特征（例如 PDH 成本、排产比例、检修损失等），说明对拐点/波动的刻画更需要“变化/形态”类特征与稳健的正则化/集成。



(a) 与 y 的相关性（Top-20，Spearman）



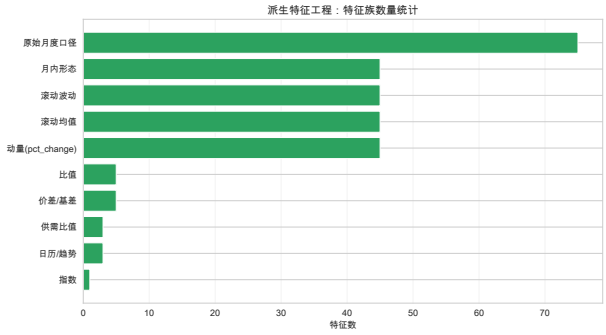
(b) 与 y_return 的相关性（Top-20，Spearman）

图 17: 特征与目标的相关性对比（pp_base；特征已按 $t-1$ 对齐）

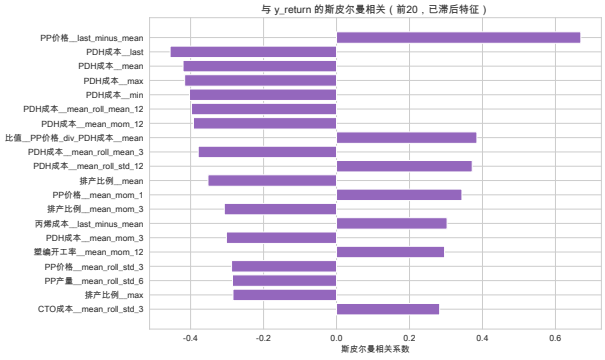
3.7 派生特征工程（可选开关）

设计原则（只用过去信息） 派生特征工程的目标是把“水平信号”转化为更稳健的“变化/形态/相对关系”信号，从而提升对拐点与阶段切换的刻画能力。为避免信息泄露，我们先完成统一月度聚合与 $\mathbf{x}_{t-1} \to y_t$ 的滞后对齐，再在已滞后序列上构造派生特征（例如滚动均值、动量、价差等），确保任何特征都不使用 t 月之后的信息。

派生特征族(与后续建模强相关) 本项目的特征工程覆盖 6 类(实现于 `pp_forecast/feature_engineering.py`)：**动量/滚动统计：**对每个因子月均值构造 `pct_change` 动量（1/3/12 月）与 3/6/12 月滚动均值、滚动波动；（2）**月内形态：**用 `range_over_mean`、`last_minus_mean` 等刻画“当月波动状态/收盘偏离”；（3）**价差/比值（利润与相对强弱）：**构造“现货-成本”（丙烯/乙烯/MTO/CTO/PDH）与“现货/成本”比值；（4）**供需比值：**进口/产量、库存/产量、检修损失/产量等相对量；（5）**下游指数：**对塑编/注塑/BOPP 开工率取均值形成需求景气综合指标；（6）**日历与趋势：**时间索引、月份正余弦以刻画季节性与长期漂移。



(a) 各特征族数量（pp_base_engineered）



(b) 与 y_return 的相关性 Top-20（pp_base_engineered）

图 18: 派生特征工程的规模与效果（示例：pp_base_engineered）

对建模的直接收益 从图 18 可以看到，派生特征显著增加了“变化/形态/相对关系”维度的信息供给：与 y_return 的相关性 Top-20 中，更多出现动量、滚动波动、月内形态与价差/比值特征，这些特征与“涨跌方向/强度”的任务定义更一致，也解释了为什么在问题 2 中“特征工程版本”的强度分类效果明显提升。与此同时，派生特征会带来维度膨胀与共线性增强，因此我们在模型侧配套使用了正则化（Ridge/ElasticNet）、稳健损失（Huber）与集成方法（RF/GBDT/Bagging/Top-k 集成）来控制过拟合风险。

4 问题 1：PP 价格月度预测与涨跌方向

4.1 问题设定与评价设计

问题 1 包含两个输出：**（1）价格点预测** $\hat{y}_t$，**（2）由点预测派生的涨跌方向** $\mathbb{I}(\widehat{y_t - y_{t-1}} > 0)$。其中方向预测严格按上一月价格 y_{t-1} 与预测价 $\hat{y}_t$ 的差值确定，保证与真实业务决策一致。

按题目提示，我们使用连续时间窗口作为测试集：默认测试窗为 **2021-01~2021-07**，训练集为其之前所有月份，并保持时间顺序（不打乱）。在训练集内部，我们额外使用时间序列交叉验证（TimeSeriesSplit）估计模型稳定性，并为集成学习提供权重依据。

评价指标包括：

- 回归：MAE、RMSE、MAPE

- 方向：Accuracy、Precision、Recall（方向定义严格按 $(y_t - y_{t-1}) > 0$ 为涨）

其中 RMSE 对极端误差更敏感，更符合价格预测中的风险偏好；方向指标用于衡量策略层面的判断能力（能否抓住涨/跌）。

4.2 基线模型（必须项）与可解释对照

基线模型仅使用历史价格序列本身，不引入外生因子，是检验“特征是否真的带来增益”的必要对照。我们实现了：

- **Naive：** $\hat{y}_t = y_{t-1}$，反映价格惯性；
- **Seasonal-12：** $\hat{y}_t = y_{t-12}$，反映年度季节性假设。

表 6 显示 naive 在测试窗上已能取得较低误差（RMSE 约 372），说明月度价格具有较强惯性；但其方向 Precision/Recall 为 0，意味着在该测试窗内 naive 的涨跌判断几乎退化为“全押同一方向”。这也从侧面说明：**要提升方向判断与拐点刻画能力，必须引入成本、库存、开工等外生信息以及变化/形态特征**。

表 6: 问题 1 基线模型（pp_base，测试窗 2021-01~2021-07）

Baseline	MAE	RMSE	MAPE(%)	Acc	Precision	Recall
naive ($\hat{y}_t = y_{t-1}$)	325.75	372.20	3.70	0.571	0.000	0.000
seasonal_12 ($\hat{y}_t = y_{t-12}$)	1336.20	1489.51	15.11	0.571	0.000	0.000

4.3 特征集与模型族

为对比“期货信息”和“派生特征工程”的边际收益，我们对四套数据集并行训练与评测：`pp_base` / `pp_base_engineered` / `pp_with_futures` / `pp_with_futures_engineered`。其中“含期货”方案由于覆盖短，默认采用 `restrict` 以保证期货变量真实可用。

模型方面，我们覆盖了线性/非线性/集成三大类，并统一在 Pipeline 内做缺失填补、缺失指示与标准化（线性/距离模型尤为关键）。主要模型与关键参数见表 7，完整参数会在运行后输出到 `outputs/metrics/<dataset>/pp_params_<model>.json`。

表 7: 主要模型与参数设置（节选）

模型	关键参数
Ridge	$\alpha = 1.0$
Lasso	$\alpha = 0.001$，max_iter=20000
ElasticNet	$\alpha = 0.001$，l1_ratio=0.5，max_iter=20000
BayesianRidge	默认参数
HuberRegressor	max_iter=2000，$\epsilon = 1.35$，$\alpha = 1e-4$
KNNRegressor	n_neighbors=8，weights=distance
SVR(RBF)	$C = 10$，γ =scale，$\epsilon = 0.1$
RandomForestRegressor	n_estimators=500，max_depth=6，min_samples_leaf=2
ExtraTreesRegressor	n_estimators=1000，max_depth=8，min_samples_leaf=2
GBR	learning_rate=0.05，n_estimators=500，max_depth=3
AdaBoostRegressor	n_estimators=500，learning_rate=0.05
Bagging(Tree)	n_estimators=300，bootstrap=True，base_tree_depth=4
LogisticRegression(强度)	max_iter=2000
SVC(RBF, 强度)	$C = 5$，γ =scale，probability=True
RandomForestClassifier(强度)	n_estimators=500，max_depth=6，min_samples_leaf=2
ExtraTreesClassifier(强度)	n_estimators=1000，max_depth=8，min_samples_leaf=2
GBR_clf(强度)	learning_rate=0.05，n_estimators=500，max_depth=3
AdaBoostClassifier(强度)	n_estimators=500，learning_rate=0.05
Bagging(Tree, 强度)	n_estimators=300，bootstrap=True，base_tree_depth=4
SARIMAX(可选)	order=(1,1,1)，季节项 (1,0,1,12)(样本足够时启用)，外生变量取 Top-20 相关特征
Prophet(可选)	yearly_seasonality=True，additive；外生回归量取 Top-10 相关特征

4.4 集成学习与模型选择策略

考虑到月度样本量有限、单模型不稳定，我们在回归侧实现了多种集成策略（均值/中位数/截尾/训练集 CV 加权/Top-k CV 加权）。其核心思想是：**把“模型不确定性”显式平均掉**，尤其在拐点月份可减少某个模型偶然偏离导致的大误差。

其中 `ensemble_cv_weighted` 会在训练集上做时间序列交叉验证，按 $1/\mathrm{RMSE}^2$ 归一化得到权重，对测试集预测做加权平均；`ensemble_topk_cv_weighted` 则只选取 CV 表现最好的 Top-k 模型，进一步抑制弱模型拖累。

4.5 结果、可视化与误差分析

表 8 汇总了四种方案在测试窗上的最优结果（按 RMSE 选择）。整体结论非常清晰：**长样本 + 特征工程（pp_base_engineered）**在 RMSE/MAE 与方向指标上均最优；而“含期货 (restrict)”方案由于训练样本显著变短，在该测试窗内反而劣于不含期货的长样本方案。

表 8: 问题 1：价格点预测与涨跌方向（最佳模型摘要，按 RMSE 选取）

数据集	方案	最佳模型	MAE	RMSE	MAPE(%)	Acc	Precision/Recall
pp_base	不含期货	ensemble_cv_weighted	192.90	220.62	2.19	0.857	0.750 / 1.000
pp_base_engineered	不含期货 + 特征工程	ensemble_cv_weighted	126.33	155.81	1.46	1.000	1.000 / 1.000
pp_with_futures	含期货 (restrict)	lasso	463.86	536.11	5.30	0.571	0.500 / 0.667
pp_with_futures_engineered	含期货 + 特征工程 (restrict)	ensemble_topk_cv_weighted	474.06	559.79	5.37	0.571	0.500 / 0.333

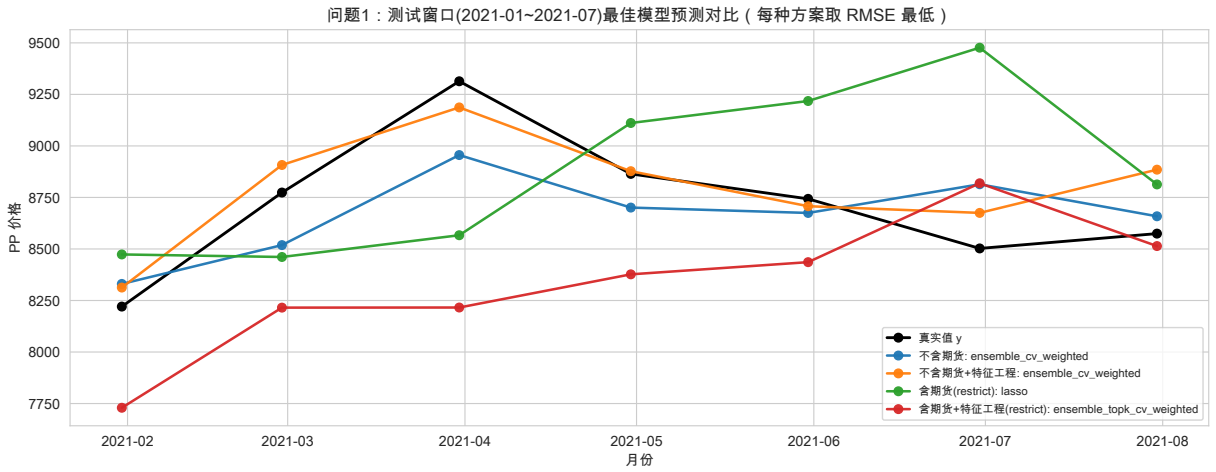


图 19: 测试窗最佳模型预测对比（每种方案取 RMSE 最低；由 `scripts/run_pp_report_figures.py` 生成）

关键讨论

- **特征工程显著降低误差：**在不含期货的长样本上，引入“动量/滚动统计/价差/供需比值/形态”后 RMSE 从约 221 降至约 156，且方向在该 7 个月测试窗上达到 1.0（需注意测试窗较短，仍需更长滚动回测检验稳定性）。
- **期货信息的样本代价：**期货变量覆盖仅约 22 个月，`restrict` 会大幅缩短训练集，导致模型更依赖填补/正则化，整体效果不如长样本；因此期货更适合作为“敏感性对照”或在未来补全数据后再纳入主模型。
- **集成学习的稳健性：**图 19 中，CV 加权集成整体更平滑、更少出现单点跳变，符合“平均模型不确定性”的设计目标；在行情快速变化月份，集成能一定程度抑制过度追涨杀跌。
- **解释与业务含义：**结合 EDA 与问题 3 的阶段筛选结果，成本（尤其丙烯/乙烯/煤化工路线）与库存通常是解释价格水平的主因子；而方向/强度更依赖变化与形态特征（动量、价差变化、月内波动状态）。

5 问题 2：涨跌强度预测与概率输出

5.1 任务设定与难点

问题 2 要求输出“涨跌强度”的离散档位及其概率。与问题 1 不同，强度预测更关注**变化幅度**而非价格水平：在 EDA 中我们已经看到 y_return 的相关性更弱、且更依赖冲击与形态特征；同时强度五分类存在类别不均衡（`flat` 较少），因此需使用 macro-F1 等更公平的指标。

5.2 两条建模路径（数值 vs 概率）

本项目同时提供两条实现路径并在输出文件中完整保留：

- **路径 A（回归派生）：**先预测 $\hat{y}_t$，再计算 $\hat{r}_t = (\hat{y}_t - y_{t-1})/y_{t-1}$，并按阈值映射到强度档位（得到**数值涨跌幅预测**与离散档位）。
- **路径 B（强度多分类）：**直接以 `y_strength` 五分类建模，输出每一档概率 `proba__*`（满足题目“输出出现相关涨跌幅度的概率”）。

两条路径的差异在于：路径 A 的概率需要额外假设（如残差分布）才能得到，而路径 B 天然输出概率但更依赖特征对“变化幅度”的解释力。为此，我们在路径 B 中重点使用派生特征工程（动量/滚动波动/价差等）并采用集成分类器增强稳定性。

5.3 评价指标与概率输出口径

对强度五分类，报告：Accuracy、macro-Precision、macro-Recall、macro-F1（时间窗同问题 1）。其中 macro-F1 对少数类更敏感，更能反映“极端涨跌”识别能力。

概率输出采用 `predict_proba`：每个测试月给出五档概率向量 $\hat{p}(\text{class} \mid \mathbf{x}_{t-1})$。此外，对于路径 A 我们提供可选 residual bootstrap（`--bootstrap N`），在不改变点预测模型的前提下，通过重采样训练残差近似得到价格区间、收益率区间以及 $p(\text{up})$。

5.4 结果摘要与可视化解释

表 9 给出四套数据集的最佳强度分类结果（按 macro-F1 选取）。可以看到：**特征工程对强度任务收益显著**，`pp_base_engineered` 的最佳模型在本次运行中取得最高 macro-F1；而含期货方案由于样本缩短，对五分类的泛化更不稳定。

表 9: 问题 2：涨跌强度（五分类）与概率输出（最佳模型摘要，按 macro-F1 选取）

数据集	方案	最佳模型	F1_macro	Acc	Precision_macro	Recall_macro
pp_base	不含期货	ensemble_proba_topk_cv_weighted	0.600	0.571	0.750	0.750
pp_base_engineered	不含期货 + 特征工程	ada_clf	0.852	0.857	0.933	0.833
pp_with_futures	含期货 (restrict)	logreg	0.354	0.571	0.438	0.312
pp_with_futures_engineered	含期货 + 特征工程 (restrict)	logreg	0.333	0.571	0.400	0.312

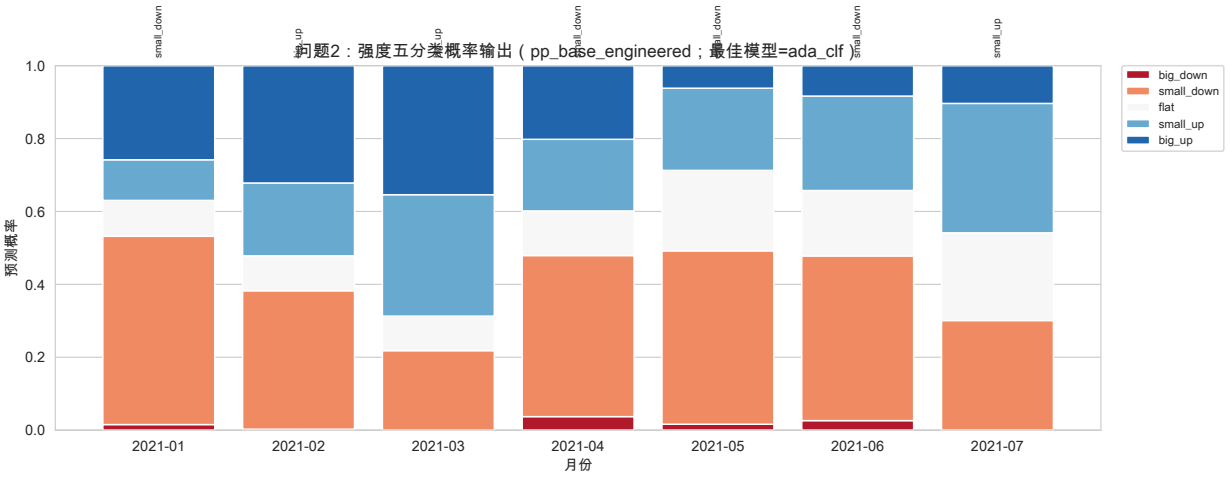


图 20: 强度五分类概率输出示例（pp_base_engineered；测试窗；由 `scripts/run_pp_report_figures.py` 生成）

概率输出文件与业务解读 对每个测试月，`pp_strength_test_predictions.csv` 与 `pp_strength_ensemble_te` 给出 `strength_pred` 以及各档概率；回归路径的数值涨跌幅预测见 `pp_test_predictions.csv` 的 `return_pred`。

业务上，概率输出可直接用于风险管理：当概率集中在 `big_up/big_down` 时，意味着模型认为行情处于“高波动偏向”状态，可对应更强的库存/套保策略；当概率主要集中于 `flat` 或相邻档位时，说明不确定性较高，更适合采用保守策略或等待更多信息。

6 问题 3：分阶段关键因子筛选与权重

6.1 动机：因子作用的阶段性

从滚动滞后相关（图 11）可以看到，成本类因子与价格的关系相对稳定，而库存、开工率、产量等“量”变量的相关方向与强度更容易随阶段翻转。这意味着用“全样本一套固定权重”去解释因子贡献往往不稳健。因此问题 3 的目标是：**先识别行业周期阶段，再在每个阶段内自动筛选关键因子组并输出权重**。

6.2 阶段划分方法（趋势 × 波动）

我们用收益率构造趋势与波动特征（默认窗口均为 6 个月）：

$$\text{trend}_t = \frac{y_t}{y_{t-6}} - 1, \qquad \text{vol}_t = \text{Std}(r_{t-5:t})$$

其中趋势按阈值(默认 ±2%)分为 `up/down/flat`，波动按分位数(默认 0.6 分位)分为 `high_vol/low_vol`，两者拼接得到 `regime`（例如 `up_high_vol`）。为避免短期噪声导致频繁切换，我们将长度小于 6 个月的 regime 段合并到相邻段（优先合并到收益率均值更接近的一侧），得到最终阶段 `stage_id`。阶段划分的输出见表 10 与图 21（由 `scripts/run_q3_stage_factor_selection.py` 生成）。

表 10: 问题 3：阶段划分结果摘要（以 pp_base 为例）

阶段	市场状态 (regime)	起止月份	月数	累计涨跌 (%)	波动 (%/月)
1	flat_high_vol	2015-01 ~ 2015-08	8	-3.22	5.49
2	down_low_vol	2015-09 ~ 2016-02	6	-19.94	3.57
3	up_high_vol	2016-03 ~ 2017-03	13	16.32	5.99
4	up_low_vol	2017-04 ~ 2018-10	19	35.27	3.24
5	down_low_vol	2018-11 ~ 2020-03	17	-31.54	3.45
6	up_low_vol	2020-04 ~ 2021-07	16	16.37	3.87

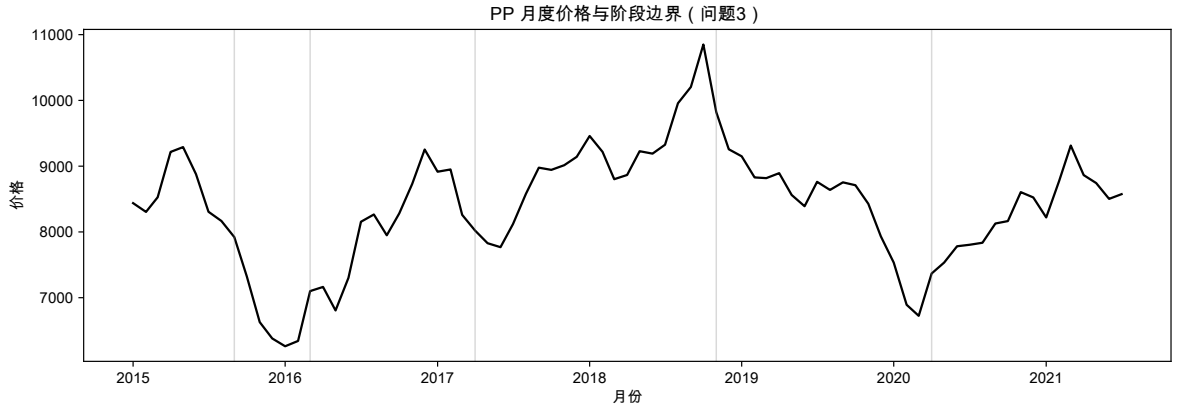


图 21: PP 月度价格与阶段边界（pp_base）

6.3 阶段内关键因子组自动筛选与权重计算

在每个阶段内，我们以 Ridge 回归拟合 $\mathbf{x}_{t-1} \to y_t$，并输出系数。为提升可解释性，我们把特征按“因子组”聚合（同一前缀，如 丙烯成本 `__mean`、丙烯成本 `__range` 等都归入 丙烯成本组），

并对组内系数绝对值求和得到组权重：

$$w(g) = \frac{\sum_{f \in g} |\beta_f|}{\sum_f |\beta_f|}$$

其中 $w(g)$ 可解释为“该阶段内该因子组对预测的相对贡献”。Ridge 的作用是抑制成本路线间的共线性导致的系数不稳定；同时我们默认不将 PP 价格自身作为候选因子组（避免自回归项掩盖外生因子），从而更聚焦于“供需/成本/库存/需求/宏观”等解释因素。

表 11 给出 pp_base 的 Top-5 因子组示例；同时，我们也对强度任务在每个阶段内训练多项 Logistic 回归并输出组权重（见 q3_strength_logreg_group_weights.csv），用于解释“哪些因子更影响涨跌幅度的分档”。

表 11: 问题 3：各阶段关键因子组（Top-5，Ridge 绝对系数归一化权重，pp_base）

阶段	regime	Top-5 因子组（权重）
1	flat_high_vol	丙烯成本 (0.177)，PP 产量 (0.154)，PP 石化库存 (0.142)，乙烯成本 (0.113)，排产比例 (0.109)
2	down_low_vol	CTO 成本 (0.161)，塑编开工率 (0.150)，丙烯成本 (0.148)，排产比例 (0.118)，PP 开工率 (0.115)
3	up_high_vol	PP 石化库存 (0.221)，乙烯成本 (0.152)，CTO 成本 (0.142)，PP 产量 (0.097)，排产比例 (0.096)
4	up_low_vol	乙烯成本 (0.150)，塑编开工率 (0.142)，PP 石化库存 (0.129)，排产比例 (0.104)，CTO 成本 (0.100)
5	down_low_vol	丙烯成本 (0.153)，CTO 成本 (0.102)，乙烯成本 (0.087)，PDH 成本 (0.084)，检修损失 (0.083)
6	up_low_vol	排产比例 (0.169)，PP 石化库存 (0.149)，PDH 成本 (0.122)，PP 开工率 (0.106)，乙烯成本 (0.078)

为体现“结合问题 2 模型”的阶段解释，我们同样对强度任务在每个阶段内训练多项 Logistic 回归，并按系数幅度归一化得到因子组权重；Top-5 结果见表 12。

表 12: 问题 3（强度任务）：各阶段关键因子组（Top-5，LogReg 系数幅度归一化权重，pp_base）

阶段	regime	Top-5 因子组（权重）
1	flat_high_vol	排产比例 (0.147)，塑编开工率 (0.124)，乙烯成本 (0.116)，PP 进口量 (0.104)，PP 石化库存 (0.102)
2	down_low_vol	排产比例 (0.130)，PP 石化库存 (0.123)，丙烯成本 (0.117)，PP 开工率 (0.101)，BOPP 开工率 (0.099)
3	up_high_vol	乙烯成本 (0.152)，排产比例 (0.140)，PP 石化库存 (0.117)，PP 开工率 (0.115)，塑编开工率 (0.110)
4	up_low_vol	PP 开工率 (0.142)，排产比例 (0.108)，塑编开工率 (0.099)，检修损失 (0.097)，MTO 成本 (0.091)
5	down_low_vol	CTO 成本 (0.136)，排产比例 (0.129)，PDH 成本 (0.093)，乙烯成本 (0.090)，MTO 成本 (0.089)
6	up_low_vol	MTO 成本 (0.116)，PP 石化库存 (0.113)，PP 开工率 (0.098)，PDH 成本 (0.095)，PP 进口量 (0.087)

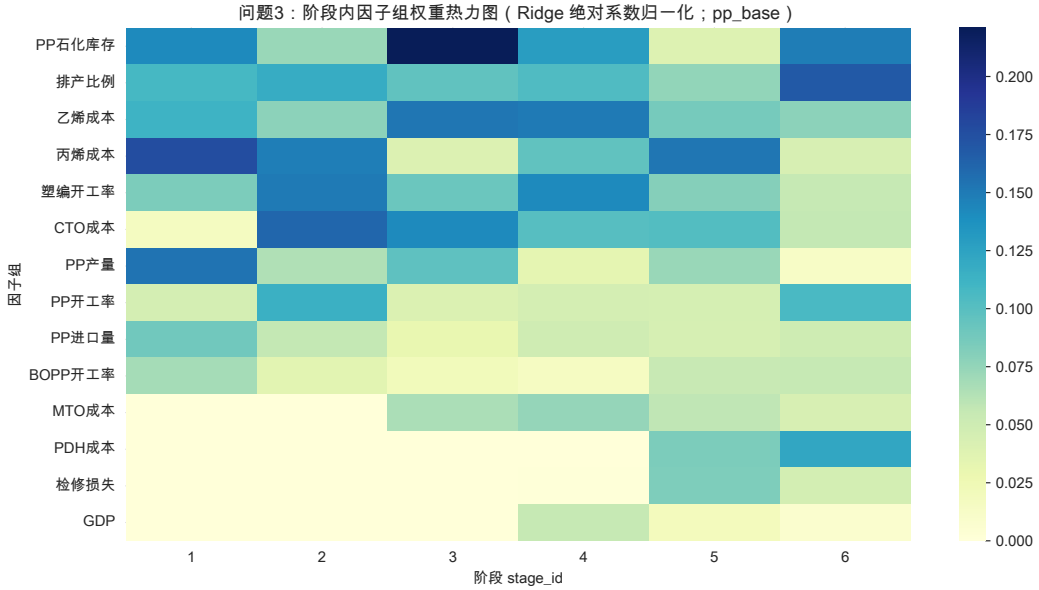


图 22: 阶段内因子组权重热力图（pp_base；Ridge 绝对系数归一化；由 scripts/run_pp_report_figures.py 生成）

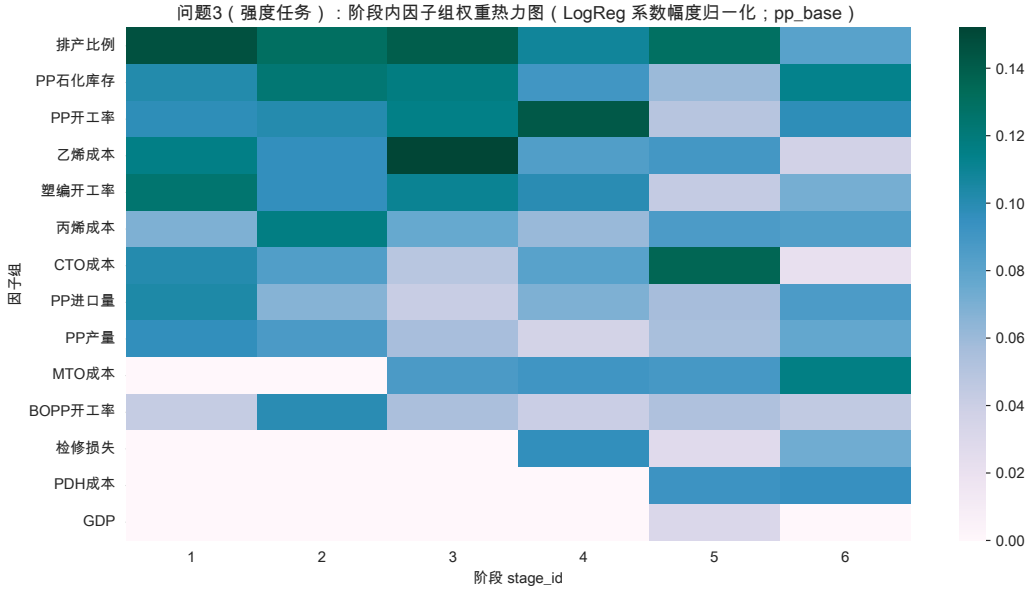


图 23: 阶段内因子组权重热力图（强度任务；LogReg 系数幅度归一化；由 `scripts/run_pp_report_figures.py` 生成）

阶段性结论与业务解释 综合表 11 与图 22，可以得到较一致的解释框架：成本类因子（丙烯/乙烯/MTO/CTO/PDH）在多数阶段都具有较高权重，代表“定价中枢”；库存与下游开工率在部分阶段显著抬升，更多反映“供需错配/风险状态”；供给侧（产量、进口、检修、排产）在高波动阶段更容易进入 Top 组，提示阶段切换往往伴随供给冲击或结构调整。GDP 作为慢变量更像背景信息，其权重通常较低但在部分阶段可作为景气解释项出现。

这类阶段化权重输出一方面能为模型提供可解释性(告诉我们“这个阶段主要在讲成本还是讲库存”)，另一方面也能反向指导特征工程：例如在高波动阶段更重视动量/波动/价差变化，在低波动阶段更重视水平与季节性。

7 总结：不足与展望

- **数据覆盖不一致：**不同因子起止时间差异显著，尤其期货覆盖较短；未来可引入“缺失机制建模”或更精细的特征筛选策略。
- **样本量有限：**月度样本数不大，复杂模型易过拟合；可进一步采用滚动回测、多窗口稳定性检验。
- **概率输出改进：**当前强度概率来自分类器 soft-voting；未来可做概率校准（Platt/Isotonic）或分布预测（分位数回归、block bootstrap）。
- **阶段划分改进：**可对比变点检测/HMM 等更贴近行业周期的分段方法，并结合业务解释阶段含义。

附录目录

A 附录 A：可复现代码结构与运行命令

Listing 1: 生成数据集 / EDA / 建模 / 问题 3（建议在 lchen 环境中运行）

```
conda run -n lchen python scripts/build_pp_dataset.py --target-metric mean --
    output outputs/datasets/pp_base.csv
conda run -n lchen python scripts/build_pp_dataset.py --include-futures --
    output outputs/datasets/pp_with_futures.csv

conda run -n lchen python scripts/build_pp_dataset.py --engineer-features --
    output outputs/datasets/pp_base_engineered.csv
conda run -n lchen python scripts/build_pp_dataset.py --include-futures --
    engineer-features --output outputs/datasets/pp_with_futures_engineered.csv

conda run -n lchen python scripts/run_pp_eda.py
conda run -n lchen python scripts/run_pp_models.py --dataset outputs/datasets/
    pp_base.csv

conda run -n lchen python scripts/run_q3_stage_factor_selection.py --dataset
    outputs/datasets/pp_base.csv --also-strength
```

B 附录 B：主要输出文件说明

- **数据集：** `outputs/datasets/pp_*.csv`
- **EDA：** `outputs/eda/`（PDF 图 + CSV 统计表，如 ACF、ADF、缺失率热力图等）
- **问题 1/2 指标：**`outputs/metrics/<dataset>/pp_model_metrics.csv`、`pp_strength_model_metrics.csv`
- **问题 2 概率：**`pp_strength_test_predictions.csv`、`pp_strength_ensemble_test_predictions.csv`
- **Bootstrap：** `pp_bootstrap_predictions.csv`（可选开关 `--bootstrap`）
- **问题 3：** `outputs/q3/<dataset>/q3_ridge_group_weights.csv`、`q3_stages_price.pdf` 等

C 附录 C：完整模型对比表（CSV 路径）

C.1 问题 1：全模型对比（回归/方向）

表 13: 问题 1：全模型回归指标明细（测试窗 2021-01~2021-07；含集成与可选时序模型）

数据集	模型	MAE	RMSE	MAPE(%)
pp_base	ensemble_cv_weighted	192.90	220.62	2.19
pp_base	ensemble_median	201.82	222.77	2.31
pp_base	ada	216.73	231.44	2.48
pp_base	rf	224.71	242.18	2.59
pp_base	bagging_tree	223.74	244.41	2.58
pp_base	ensemble_topk_mean	225.72	246.09	2.60
pp_base	ensemble_topk_cv_weighted	225.64	246.56	2.60
pp_base	extra_trees	206.65	252.87	2.36
pp_base	ensemble_trimmed_mean	222.38	256.12	2.52
pp_base	ensemble_mean	228.46	265.30	2.61
pp_base	gbr	305.13	337.76	3.55
pp_base	svr_rbf	309.85	393.97	3.48
pp_base	bayes_ridge	338.33	419.10	3.84
pp_base	knn	391.50	510.91	4.37
pp_base	ridge	859.03	907.46	9.84
pp_base	lasso	806.24	981.16	9.29
pp_base	elasticnet	812.01	1023.55	9.40
pp_base	huber	1166.52	1426.14	13.29
pp_base	sarimax	1773.90	1822.10	20.22
pp_base_engineered	ensemble_cv_weighted	126.33	155.81	1.46
pp_base_engineered	ensemble_trimmed_mean	137.38	175.32	1.58
pp_base_engineered	ensemble_median	162.02	181.65	1.86
pp_base_engineered	ensemble_topk_mean	158.53	189.36	1.82
pp_base_engineered	ensemble_topk_cv_weighted	163.05	193.50	1.87
pp_base_engineered	ensemble_mean	146.79	218.87	1.70
pp_base_engineered	bayes_ridge	203.04	221.45	2.34
pp_base_engineered	ada	189.46	225.20	2.15
pp_base_engineered	gbr	225.71	239.59	2.60
pp_base_engineered	rf	230.82	243.11	2.66
pp_base_engineered	bagging_tree	229.28	243.18	2.64
pp_base_engineered	extra_trees	202.89	261.02	2.31
pp_base_engineered	svr_rbf	308.21	392.64	3.46
pp_base_engineered	knn	325.04	418.68	3.65
pp_base_engineered	sarimax	479.25	599.53	5.39
pp_base_engineered	ridge	584.68	644.58	6.72
pp_base_engineered	huber	630.34	720.52	7.24
pp_base_engineered	lasso	665.65	952.89	7.66
pp_base_engineered	elasticnet	1008.26	1278.00	11.46
pp_with_futures	lasso	463.86	536.11	5.30
pp_with_futures	knn	773.96	835.77	8.76
pp_with_futures	ada	730.59	862.87	8.22
pp_with_futures	bagging_tree	763.39	867.42	8.61
pp_with_futures	ensemble_topk_mean	711.47	890.49	8.01
pp_with_futures	ensemble_topk_cv_weighted	715.98	893.44	8.06
pp_with_futures	rf	792.46	895.09	8.94

续下页

表 13: 问题 1：全模型回归指标明细（续；测试窗 2021-01~2021-07）

数据集	模型	MAE	RMSE	MAPE(%)
pp_with_futures	ensemble_mean	783.80	901.55	8.84
pp_with_futures	ensemble_cv_weighted	771.82	903.50	8.70
pp_with_futures	ensemble_trimmed_mean	835.22	943.77	9.43
pp_with_futures	ensemble_median	834.50	948.53	9.42
pp_with_futures	svr_rbf	900.71	954.19	10.22
pp_with_futures	extra_trees	871.23	989.16	9.83
pp_with_futures	huber	869.03	1008.92	9.82
pp_with_futures	bayes_ridge	869.03	1008.92	9.82
pp_with_futures	elasticnet	866.63	1074.01	9.76
pp_with_futures	gbr	1015.85	1138.89	11.50
pp_with_futures	sarimax	896.51	1197.12	10.13
pp_with_futures	ridge	1113.73	1234.88	12.62
pp_with_futures_engineered	ensemble_topk_cv_weighted	474.06	559.79	5.37
pp_with_futures_engineered	lasso	450.08	565.37	5.20
pp_with_futures_engineered	ensemble_topk_mean	486.06	571.17	5.50
pp_with_futures_engineered	ensemble_cv_weighted	561.48	670.64	6.34
pp_with_futures_engineered	ensemble_mean	599.50	719.12	6.76
pp_with_futures_engineered	bayes_ridge	624.08	747.66	7.07
pp_with_futures_engineered	huber	624.08	747.66	7.07
pp_with_futures_engineered	ridge	639.20	765.30	7.24
pp_with_futures_engineered	ensemble_trimmed_mean	660.26	771.38	7.45
pp_with_futures_engineered	knn	757.52	818.31	8.58
pp_with_futures_engineered	ensemble_median	748.00	826.64	8.46
pp_with_futures_engineered	ada	765.76	870.20	8.63
pp_with_futures_engineered	bagging_tree	792.33	875.09	8.95
pp_with_futures_engineered	rf	817.26	895.01	9.24
pp_with_futures_engineered	gbr	844.59	917.66	9.56
pp_with_futures_engineered	svr_rbf	900.69	954.16	10.22

续下页

表 13: 问题 1：全模型回归指标明细（续；测试窗 2021-01~2021-07）

数据集	模型	MAE	RMSE	MAPE(%)
pp_with_futures_engineered	extra_trees	1002.17	1126.38	11.33
pp_with_futures_engineered	elasticnet	1244.56	1328.76	14.40
pp_with_futures_engineered	sarimax	11049.62	12389.01	126.24

表 14: 问题 1：全模型方向指标明细（测试窗 2021-01~2021-07；由回归点预测派生方向）

数据集	模型	Acc	Precision	Recall
pp_base	ensemble_cv_weighted	0.857	0.750	1.000
pp_base	ensemble_median	0.857	0.750	1.000
pp_base	ada	0.857	0.750	1.000
pp_base	rf	0.857	0.750	1.000
pp_base	bagging_tree	0.857	0.750	1.000
pp_base	ensemble_topk_mean	0.714	0.600	1.000
pp_base	ensemble_topk_cv_weighted	0.714	0.600	1.000
pp_base	extra_trees	0.857	0.750	1.000
pp_base	ensemble_trimmed_mean	1.000	1.000	1.000
pp_base	ensemble_mean	0.714	0.667	0.667
pp_base	gbr	0.571	0.500	1.000
pp_base	svr_rbf	0.714	1.000	0.333
pp_base	bayes_ridge	0.571	0.500	0.333
pp_base	knn	0.714	1.000	0.333
pp_base	ridge	0.571	0.000	0.000
pp_base	lasso	0.857	1.000	0.667
pp_base	elasticnet	0.857	1.000	0.667
pp_base	huber	0.571	0.500	1.000
pp_base	sarimax	0.571	0.000	0.000
pp_base_engineered	ensemble_cv_weighted	1.000	1.000	1.000
pp_base_engineered	ensemble_trimmed_mean	1.000	1.000	1.000
pp_base_engineered	ensemble_median	1.000	1.000	1.000
pp_base_engineered	ensemble_topk_mean	0.857	0.750	1.000
pp_base_engineered	ensemble_topk_cv_weighted	0.857	0.750	1.000
pp_base_engineered	ensemble_mean	1.000	1.000	1.000
pp_base_engineered	bayes_ridge	1.000	1.000	1.000
pp_base_engineered	ada	1.000	1.000	1.000
pp_base_engineered	gbr	0.857	0.750	1.000
pp_base_engineered	rf	0.714	0.600	1.000
pp_base_engineered	bagging_tree	0.714	0.600	1.000
pp_base_engineered	extra_trees	1.000	1.000	1.000
pp_base_engineered	svr_rbf	0.714	1.000	0.333
pp_base_engineered	knn	0.714	1.000	0.333
pp_base_engineered	sarimax	0.571	0.500	0.333

续下页

表 14: 问题 1：全模型方向指标明细（续；测试窗 2021-01~2021-07）

数据集	模型	Acc	Precision	Recall
pp_base_engineered	ridge	1.000	1.000	1.000
pp_base_engineered	huber	1.000	1.000	1.000
pp_base_engineered	lasso	0.714	0.600	1.000
pp_base_engineered	elasticnet	0.429	0.429	1.000
pp_with_futures	lasso	0.571	0.500	0.667
pp_with_futures	knn	0.571	0.000	0.000
pp_with_futures	ada	0.571	0.000	0.000
pp_with_futures	bagging_tree	0.571	0.000	0.000
pp_with_futures	ensemble_topk_mean	0.571	0.000	0.000
pp_with_futures	ensemble_topk_cv_weighted	0.571	0.000	0.000
pp_with_futures	rf	0.571	0.000	0.000
pp_with_futures	ensemble_mean	0.571	0.000	0.000
pp_with_futures	ensemble_cv_weighted	0.571	0.000	0.000
pp_with_futures	ensemble_trimmed_mean	0.571	0.000	0.000
pp_with_futures	ensemble_median	0.571	0.000	0.000
pp_with_futures	svr_rbf	0.571	0.000	0.000
pp_with_futures	extra_trees	0.571	0.000	0.000
pp_with_futures	huber	0.571	0.000	0.000
pp_with_futures	bayes_ridge	0.571	0.000	0.000
pp_with_futures	elasticnet	0.571	0.000	0.000
pp_with_futures	gbr	0.571	0.000	0.000
pp_with_futures	sarimax	0.714	1.000	0.333
pp_with_futures	ridge	0.571	0.000	0.000
pp_with_futures_ engineered	ensemble_topk_cv_weighted	0.571	0.500	0.333
pp_with_futures_ engineered	lasso	0.571	0.500	1.000
pp_with_futures_ engineered	ensemble_topk_mean	0.429	0.000	0.000
pp_with_futures_ engineered	ensemble_cv_weighted	0.571	0.000	0.000
pp_with_futures_ engineered	ensemble_mean	0.571	0.000	0.000
pp_with_futures_ engineered	bayes_ridge	0.714	1.000	0.333
pp_with_futures_ engineered	huber	0.714	1.000	0.333
pp_with_futures_ engineered	ridge	0.714	1.000	0.333
pp_with_futures_ engineered	ensemble_trimmed_mean	0.571	0.000	0.000
pp_with_futures_ engineered	knn	0.571	0.000	0.000
pp_with_futures_ engineered	ensemble_median	0.571	0.000	0.000

续下页

表 14: 问题 1：全模型方向指标明细（续；测试窗 2021-01~2021-07）

数据集	模型	Acc	Precision	Recall
pp_with_futures_engineered	ada	0.571	0.000	0.000
pp_with_futures_engineered	bagging_tree	0.571	0.000	0.000
pp_with_futures_engineered	rf	0.571	0.000	0.000
pp_with_futures_engineered	gbr	0.571	0.000	0.000
pp_with_futures_engineered	svr_rbf	0.571	0.000	0.000
pp_with_futures_engineered	extra_trees	0.571	0.000	0.000
pp_with_futures_engineered	elasticnet	0.429	0.333	0.333
pp_with_futures_engineered	sarimax	0.429	0.400	0.667

C.2 问题 2：全模型对比（强度五分类）

表 15: 问题 2：全模型强度五分类指标明细（测试窗 2021-01~2021-07；macro-F1 降序）

数据集	模型	Acc	Prec_m	Rec_m	F1_m
pp_base	ensemble_proba_topk_cv_weighted	0.571	0.750	0.750	0.600
pp_base	rf_clf	0.571	0.367	0.583	0.444
pp_base	gbr_clf	0.429	0.306	0.667	0.400
pp_base	ensemble_proba_mean	0.429	0.583	0.438	0.392
pp_base	bagging_tree_clf	0.429	0.292	0.250	0.268
pp_base	ada_clf	0.286	0.500	0.188	0.267
pp_base	knn_clf	0.571	0.190	0.333	0.242
pp_base	extra_trees_clf	0.429	0.188	0.188	0.188
pp_base	nb	0.143	0.062	0.250	0.100
pp_base	logreg	0.000	0.000	0.000	0.000
pp_base	svc_rbf	0.000	0.000	0.000	0.000
pp_base_engineered	ada_clf	0.857	0.933	0.833	0.852
pp_base_engineered	ensemble_proba_topk_cv_weighted	0.571	0.367	0.583	0.444
pp_base_engineered	bagging_tree_clf	0.429	0.583	0.583	0.433
pp_base_engineered	gbr_clf	0.429	0.375	0.438	0.375
pp_base_engineered	knn_clf	0.571	0.190	0.333	0.242
pp_base_engineered	svc_rbf	0.571	0.190	0.333	0.242
pp_base_engineered	extra_trees_clf	0.429	0.167	0.250	0.200
pp_base_engineered	logreg	0.429	0.188	0.188	0.188
pp_base_engineered	ensemble_proba_mean	0.286	0.100	0.125	0.111
pp_base_engineered	rf_clf	0.143	0.056	0.333	0.095
pp_base_engineered	nb	0.143	0.083	0.062	0.071

续下页

表 15: 问题 2：全模型强度五分类指标明细（续；测试窗 2021-01~2021-07）

数据集	模型	Acc	Prec_m	Rec_m	F1_m
pp_with_futures	logreg	0.571	0.438	0.312	0.354
pp_with_futures	ensemble_proba_topk_cv_weighted	0.571	0.400	0.312	0.333
pp_with_futures	nb	0.571	0.190	0.333	0.242
pp_with_futures	svc_rbf	0.571	0.190	0.333	0.242
pp_with_futures	extra_trees_clf	0.571	0.200	0.250	0.222
pp_with_futures	rf_clf	0.571	0.167	0.250	0.200
pp_with_futures	bagging_tree_clf	0.429	0.125	0.188	0.150
pp_with_futures	ensemble_proba_mean	0.429	0.125	0.188	0.150
pp_with_futures	gbr_clf	0.143	0.083	0.062	0.071
pp_with_futures	ada_clf	0.143	0.083	0.062	0.071
pp_with_futures	knn_clf	0.000	0.000	0.000	0.000
pp_with_futures_engineered	logreg	0.571	0.400	0.312	0.333
pp_with_futures_engineered	svc_rbf	0.571	0.190	0.333	0.242
pp_with_futures_engineered	ensemble_proba_mean	0.571	0.190	0.333	0.242
pp_with_futures_engineered	rf_clf	0.571	0.167	0.250	0.200
pp_with_futures_engineered	bagging_tree_clf	0.571	0.167	0.250	0.200
pp_with_futures_engineered	nb	0.429	0.150	0.188	0.167
pp_with_futures_engineered	extra_trees_clf	0.429	0.150	0.188	0.167
pp_with_futures_engineered	ensemble_proba_topk_cv_weighted	0.429	0.150	0.188	0.167
pp_with_futures_engineered	ada_clf	0.143	0.125	0.062	0.083
pp_with_futures_engineered	gbr_clf	0.143	0.083	0.062	0.071
pp_with_futures_engineered	knn_clf	0.000	0.000	0.000	0.000

CSV 输出路径（可复现与二次分析） 为便于二次分析与复现实验，完整模型对比明细同时以 CSV 形式保存于：

- `outputs/metrics/pp_base/pp_model_metrics.csv`
- `outputs/metrics/pp_base/pp_strength_model_metrics.csv`
- 其余数据集同理（目录名为数据集 stem）